

Agenti, Causalità e Allineamento: JEPA come fondamento per un Intelligenza Artificiale verificabile

Abstract

La presente monografia costituisce il terzo pannello di un trittico dedicato alle architetture Joint Embedding Predictive (JEPA) e al loro ruolo nel percorso verso una forma di intelligenza artificiale strutturalmente radicata. Mentre la prima monografia — *Rappresentazione, previsione e mondo interno* — ha indagato il piano percettivo-architetturale e quello filosofico-critico, e la seconda — *Var-T-JEPA e l'estensione linguistica* — ha sviluppato la formalizzazione probabilistica e l'applicazione ai modelli linguistici di grande scala, il presente lavoro abita un territorio che entrambe le opere attraversavano senza sostarvi: il piano agentivo, causale e normativo.

Tre programmi di ricerca ne formano il nucleo. Il primo è il Causal Representation Learning applicato alle architetture JEPA: si mostra come il grafo aciclico diretto implicito nella struttura generativa di Var-JEPA possa essere reso esplicito e disciplinato attraverso gli strumenti del do-calcolo di Pearl, e come i modelli Causal VAE, iVAE e i framework neuro-simbolici di terza generazione forniscano le primitive architetture necessarie a identificare fattori latenti causalmente interpretabili. Il secondo programma riguarda i sistemi Vision-Language-Action — RT-2, PaLM-E, OpenVLA e le architetture che ne discendono — come istanziazione del ciclo percezione-azione-apprendimento in ambienti fisici, con JEPA nel ruolo di modulo percettivo privilegiato; l'analisi si estende ai World Models gerarchici per la pianificazione a lungo orizzonte e al problema fondamentale del sim-to-real gap come sfida computazionale ed epistemica irrisolta. Il terzo programma abbraccia l'interpretabilità meccanicistica e l'allineamento: le tecniche di probing causale, TCAV e Sparse Autoencoder vengono applicate agli spazi latenti JEPA per rendere verificabile dall'esterno ciò che questi spazi rappresentano; il framework RLHF e il scalable oversight vengono poi analizzati come meccanismi di ancoraggio normativo di sistemi AMI.

La tesi unificante è che l'interventional distribution di Pearl e il DAG generativo di Var-JEPA non siano concetti paralleli ma strutturalmente connessi: esplicitare questa connessione non è soltanto un esercizio di unificazione formale, ma produce conseguenze architetture verificabili — sistemi che non soltanto predicono ma intervengono, non soltanto apprendono ma si allineano. La

monografia mappa le promesse di questo programma, ne deriva le conseguenze matematiche dove possibile, e segnala con chiarezza le lacune dove la letteratura disponibile non è ancora abbastanza matura da supportare affermazioni dimostrative.

Introduzione

Il territorio non occupato

Ogni sistema teorico sufficientemente ambizioso proietta ombre sui propri lati. Le due monografie precedenti hanno costruito un edificio imponente: la prima ha fissato i fondamenti percettivi e filosofici del paradigma JEPA, ha mostrato come l'architettura I-JEPA di Assran et al., le sue estensioni visive con V-JEPA, il framework AMI di LeCun e la tradizione dei World Models da Schmidhuber a Hafner appartengano a un'unica corrente intellettuale che risponde alla domanda «cos'è una buona rappresentazione»; la seconda ha addentrato questa risposta nel dominio probabilistico e linguistico, derivando con rigore il framework Var-JEPA di Gögl e Yau, la sua istanziazione tabulare, e l'estensione LLM-JEPA di Huang, LeCun e Balestriero. Ma entrambe le opere, per costruzione e scelta deliberata, hanno lasciato senza trattazione sistematica tre territori che non sono residuali rispetto al progetto JEPA — ne sono, in un senso preciso, la frontiera naturale.

Il primo è il problema della causalità come architettura, non come aspirazione. Le due monografie precedenti invocano ripetutamente la causalità come criterio di valutazione — le buone rappresentazioni dovrebbero catturare strutture causali, i World Models dovrebbero modellare relazioni di dipendenza non soltanto correlazionali — ma non derivano mai formalmente come si costruisce un modello causale sui fattori latenti di JEPA, né come si integra il formalismo interventistico di Pearl nel training objective. Questa lacuna non è di natura terminologica: è una lacuna architeturale precisa, che separa sistemi che *parlano* di causalità da sistemi che *computano* distribuzioni interventali.

Il secondo è il ciclo percezione-azione come programma di ricerca compiuto. V-JEPA 2, menzionato in entrambe le monografie, dimostra che un modello addestrato con obiettivi JEPA sul dominio video sviluppa una comprensione del movimento e della fisica qualitativa sufficiente a guidare la pianificazione — ma questo risultato viene citato, non analizzato nella sua architettura concreta. I sistemi Vision-Language-Action contemporanei — RT-2 di Google DeepMind, PaLM-E di Google Research, OpenVLA — rappresentano il tentativo più serio di chiudere il ciclo tra percezione e azione in ambienti fisici, e la domanda di come JEPA si inserisca in questa architettura non è stata posta con la precisione necessaria. Il sim-to-real gap — la discrepanza sistematica tra le rappresentazioni apprese in simulazione e le distribuzioni osservate nel mondo

fisico — è il problema tecnico più concreto che emerge da questo territorio, e nessuna delle due monografie lo tratta come oggetto di analisi autonomo.

Il terzo è l'interpretabilità e l'allineamento come problemi tecnici con struttura propria. Che cosa rappresentano i vettori nello spazio latente di un sistema JEPA? La domanda non è retorica: le tecniche di interpretabilità meccanicistica — probing classifiers, Testing with Concept Activation Vectors (TCAV), Sparse Autoencoders (SAE) — offrono strumenti operativi per risponderle, e i risultati di LeJEPA sulla distribuzione ottimale degli embedding forniscono il quadro teorico entro cui queste tecniche si inseriscono naturalmente. Ma l'interpretabilità è soltanto il presupposto dell'allineamento: sapere cosa un sistema rappresenta è la condizione necessaria — non sufficiente — per garantire che ciò che rappresenta corrisponda ai valori che vogliamo vengano codificati. Il Reinforcement Learning from Human Feedback (RLHF) e il scalable oversight sono le risposte operative più mature a questa sfida, ma la loro connessione con l'architettura JEPA specifica non è stata mai esplicitata.

La presente monografia occupa questi tre territori. Non è una somma dei campi — causalità, robotica, allineamento — ma un'indagine su come il paradigma JEPA, nelle sue incarnazioni canoniche e nelle sue estensioni probabilistiche, possa servire da fondamento unificante per sistemi che agiscono, inferiscono causalmente e si lasciano ispezionare e correggere. La tesi non è che JEPA risolva questi problemi — sarebbe un'affermazione ingiustificata — ma che la struttura matematica di JEPA, e in particolare della sua reinterpretazione variazionale di Var-JEPA, contenga già gli agganci teorici attraverso cui questi problemi possono essere aggrediti in modo principiato.

La struttura dell'argomentazione

La monografia si articola in cinque parti e venti capitoli. La prima parte — Capitoli da 1 a 4 — costruisce le fondamenta del Causal Representation Learning come disciplina con struttura propria, ne traccia la genealogia dagli SCM di Pearl all'IVAE di Khemakhem et al. e al CausalVAE di Yang et al., e mostra come il DAG generativo implicito in Var-JEPA possa essere reso esplicito e disciplinato attraverso il do-calcolo; il Capitolo 4 completa questa parte analizzando i framework neuro-simbolici di terza generazione — in particolare NS-CL di Mao et al. — come istanziazioni dell'architettura che emerge dalla congiunzione tra inferenza causale e spazio latente strutturato.

La seconda parte — Capitoli da 5 a 8 — affronta il ciclo percezione-azione in ambienti fisici. Il Capitolo 5 ricostruisce il ciclo teorico dalla percezione all'azione nella cornice AMI, mostrando come il Perception Module di LeCun sia già, implicitamente, un sistema JEPA orientato all'azione. I Capitoli 6 e 7 analizzano le architetture VLA — RT-2, PaLM-E e OpenVLA — con il grado di formalità necessario a valutare dove e come JEPA può sostituire o integrare i moduli percettivi standard. Il Capitolo 8 è interamente dedicato al sim-to-real gap come problema fondamentale dell'embodiment computazionale: la sua

struttura matematica, le cause teoriche della discrepanza tra distribuzione simulata e distribuzione reale, e le strategie di riduzione più promettenti.

La terza parte — Capitoli da 9 a 11 — tratta i World Models gerarchici per la pianificazione a lungo orizzonte con JEPA come modulo percettivo. Il Capitolo 9 sviluppa la formalizzazione del planning come inferenza in un modello generativo strutturato a livelli temporali multipli. Il Capitolo 10 analizza il Model Predictive Control come meccanismo di pianificazione online compatibile con rappresentazioni latenti JEPA, derivandone il ciclo predizione-ottimizzazione-azione con piena formalità matematica. Il Capitolo 11 estende l'analisi al Reinforcement Learning gerarchico, mostrando come la separazione tra Perception Module e World Model nella cornice AMI permetta di addestrare separatamente le due componenti e di comporre le policy apprese a livelli di astrazione multipli.

La quarta parte — Capitoli da 12 a 16 — sviluppa il programma dell'interpretabilità meccanicistica e dell'allineamento. Il Capitolo 12 costruisce il quadro teorico del probing causale negli spazi latenti JEPA, derivando le condizioni sotto cui un probing classifier rispecchia struttura causale piuttosto che correlazione superficiale. Il Capitolo 13 applica TCAV alle rappresentazioni JEPA, mostrando come i Concept Activation Vectors possano essere calcolati sullo spazio latente strutturato e come questo modifichi l'interpretazione dei risultati. Il Capitolo 14 tratta gli Sparse Autoencoders come strumento di decomposizione degli spazi JEPA in direzioni interpretabili, con connessioni alla teoria della rappresentazione sparsificata e alla decomposizione della varianza spiegata. Il Capitolo 15 formalizza RLHF come problema di ottimizzazione vincolata e ne analizza la compatibilità con i training objective JEPA. Il Capitolo 16 affronta il scalable oversight come problema fondamentale di allineamento dei sistemi AMI, identificando i punti di contatto e le tensioni con il paradigma JEPA.

La quinta parte — Capitoli da 17 a 20 — opera la sintesi. Il Capitolo 17 formalizza il nucleo teorico unificante: la connessione tra distribuzioni interventali di Pearl e DAG generativo di Var-JEPA come programma di ricerca con conseguenze architetture verificabili. Il Capitolo 18 discute l'intelligenza incarnata e verificabile come ideale regolativo del trittico JEPA, distinguendo ciò che è già raggiungibile da ciò che rimane speculativo. Il Capitolo 19 inventaria i limiti strutturali residui — dal problema dell'identità numerica dei concetti attraverso le modalità percettive all'impossibilità di garantire l'allineamento senza supervisione esaustiva — e li colloca nel quadro dei teoremi di impossibilità già discussi in Rapp-JEPA. Il Capitolo 20 propone un programma di ricerca futuro che emerge direttamente dalle lacune dimostrate.

Una nota epistemologica

Le due monografie precedenti hanno già avvertito il lettore sul tipo di impresa intellettuale in cui si trovano coinvolti. Vale la pena rinnovare quell'avvertenza

con una precisazione specifica per il territorio che questa terza monografia abita.

Il Causal Representation Learning, l'apprendimento agentivo e l'allineamento sono campi in cui la letteratura è più frammentata, i paper più recenti, e il rischio di confondere il desideratum con la soluzione è più alto che in qualsiasi altro settore dell'intelligenza artificiale. Questo rischio è aggravato dalla tendenza — pervasiva nella letteratura ma particolarmente acuta in questo territorio — a usare il vocabolario della causalità in senso metaforico: si dice che un sistema «comprende le relazioni causali» quando ciò che si intende è che le sue prestazioni su certi benchmark migliorano quando si privilegiano certi tipi di generalizzazione. Questa ambiguità non è innocua. Confondere la performance su un task di generalizzazione out-of-distribution con la presenza di una struttura causale genuinamente computata non è soltanto un errore terminologico: è un errore che orienta male le aspettative e produce previsioni sbagliate su dove si trovano i limiti reali dei sistemi.

La presente monografia tiene distinti, con la massima cura, tre livelli di affermazione. Il primo è il livello delle derivazioni formali: proposizioni che seguono logicamente da definizioni e assiomi condivisi, indipendentemente dall'evidenza empirica. Il secondo è il livello delle connessioni teoricamente motivate: proposizioni che non sono deducibili con rigore formale, ma che la struttura dei modelli rende plausibili e che potrebbero in linea di principio essere falsificate. Il terzo è il livello delle speculazioni strutturate: direzioni di ricerca che la presente analisi individua come promettenti senza che la letteratura disponibile fornisca evidenza sufficiente a trattarle come risultati. Laddove un'affermazione appartiene al secondo o terzo livello, lo si dice esplicitamente.

Il nucleo teorico: perché Pearl e Var-JEPA convergono

C'è una ragione profonda per cui la congiunzione tra il programma causale di Pearl e la struttura variazionale di Var-JEPA non è arbitraria. Gögl e Yau, nel dimostrare che l'architettura JEPA canonica è il limite deterministico di un modello generativo a variabili latenti accoppiate, hanno implicitamente mostrato che il training di un sistema JEPA è equivalente all'inferenza approssimata in un modello che ha una struttura probabilistica precisa. Ma quella struttura probabilistica — un prior sul latente, un encoder variazionale, un predittore che mappa uno spazio latente in un altro — è esattamente il tipo di struttura su cui Pearl e la tradizione del Causal Representation Learning chiedono di imporre vincoli causali.

La differenza tra Var-JEPA e un Causal VAE non è, a questo livello, una differenza di specie ma di completezza della specificazione: Var-JEPA definisce come z_c e z_t siano connessi attraverso un predittore s_ψ , e impone che questa connessione sia regolarizzata attraverso l'ELBO a cinque termini; un Causal VAE impone in aggiunta che i fattori latenti z siano connessi attraverso

un DAG con equazioni strutturali identificabili. L'integrazione naturale è quella di aggiungere al training objective di Var-JEPA un termine di regolarizzazione che imponga la struttura causale sul joint (z_c, z_t) , e di interpretare il predittore s_ψ non come una rete neurale arbitraria ma come il meccanismo di stima delle distribuzioni interventali condizionate al contesto.

Questa integrazione è il nucleo teorico della prima parte della monografia. Il suo sviluppo completo — dalla derivazione formale dell'ELBO causale esteso alle condizioni di identificabilità che ne garantiscono la non-trivialità — occupa i Capitoli da 1 a 4 e costituisce il contributo teorico più originale dell'opera. Il lettore interessato soltanto alla parte agentiva o a quella dell'allineamento può seguire le parti successive autonomamente, tornando ai Capitoli 1–4 ogni volta che le parti successive invocano nozioni di causalità latente.

Sul destinatario presupposto

Questo testo presuppone che il lettore abbia già letto — o sia disposto a leggere — le due monografie precedenti. Non per ragioni di forma, ma perché le fondamenta teoriche e le notazioni su cui questo lavoro si appoggia sono costruite lì: la derivazione dell'ELBO a cinque termini di Var-JEPA, il framework AMI di LeCun con i suoi sei moduli, la geometria degli spazi JEPA derivata da LeJEPA, la critica dei teoremi di impossibilità. Richiamare ogni concetto da capo allungherebbe il testo senza aggiungere informazione. Dove un richiamo è indispensabile per seguire la derivazione, si fornisce un rimando preciso al capitolo o alla sezione della monografia gemella.

Sul piano dello stile: come nelle opere precedenti, le derivazioni matematiche non sono separate dalla narrazione ma ne fanno parte integrante. Il simbolismo Unicode è preferito alla notazione LaTeX non per ragioni estetiche ma perché rende il testo leggibile anche fuori da un ambiente di typesetting specializzato. La notazione è introdotta alla prima occorrenza e poi usata sistematicamente; le eccezioni — quando lo stesso simbolo assume significati diversi in contesti diversi — sono segnalate esplicitamente.

Una dichiarazione finale sull'onestà intellettuale. Questo campo — la connessione tra JEPA, causalità, embodiment e allineamento — è uno di quelli in cui la distanza tra il programma di ricerca e i risultati consolidati è massima. Sarebbe facile, e scorretto, presentare questo divario come già colmato. La monografia non lo fa: ogni capitolo distingue tra ciò che si può derivare, ciò che è empiricamente supportato, e ciò che è ancora una scommessa teorica. Questo non è pessimismo — è il rispetto che un programma di ricerca serio merita.

Capitolo 1 — Il framework causale di Pearl e il problema dell'apprendimento delle rappresentazioni

1.1 La gerarchia causale e la sua irriducibilità statistica

Nell'estate del 1995, in un seminario alla UCLA, Judea Pearl propose quello che sarebbe diventato uno degli argomenti più citati e più fraintesi della storia recente dell'intelligenza artificiale: l'affermazione che nessuna quantità ricavabile da una distribuzione di probabilità puramente osservativa può, da sola, determinare l'effetto di un'azione. Questa affermazione, che Pearl avrebbe poi formalizzato come il *teorema di non-identificabilità* delle distribuzioni interventali a partire dai dati osservativi in presenza di confondenti non osservati, non è un risultato negativo nel senso di una limitazione contingente: è la dimostrazione che esiste una discontinuità di tipo logico — non soltanto computazionale — tra il mondo dell'associazione statistica e il mondo della causalità.

La struttura di questa discontinuità è catturata dalla *gerarchia causale* (Pearl 2000, 2009), articolata in tre livelli che non si riducono l'uno all'altro. Al livello più basso si trova l'associazione: la capacità di rispondere a domande della forma «osservando $X = x$, quanto è probabile che $Y = y$?», formalmente $P(Y = y \mid X = x)$. Qualsiasi sistema di apprendimento che ottimizza una funzione di perdita su dati i.i.d. opera a questo livello. Al livello intermedio si trova l'intervento: la capacità di rispondere a domande della forma «se forzo $X = x$ (indipendentemente da ciò che lo avrebbe naturalmente causato), quanto è probabile che $Y = y$?», formalmente $P(Y = y \mid \text{do}(X = x))$. Al livello più alto si trova il controfattuale: la capacità di rispondere a domande della forma «sapendo che $X = x$ e $Y = y$ si sono verificati, cosa sarebbe successo a Y se X fosse stato x' invece?», formalmente $P(Y_{\{x'\}}(u) = y' \mid X = x, Y = y)$, dove il pedice x' indica la potenziale risposta — la variabile Y valutata nel mondo in cui X è stato forzato a x' .

Questa gerarchia è irriducibile nel senso preciso che non esiste trasformazione algebrica che permetta di ricavare le quantità di livello superiore da quelle di livello inferiore senza assunzioni aggiuntive. La conoscenza di $P(Y \mid X)$ non determina $P(Y \mid \text{do}(X))$ in generale; la conoscenza di $P(Y \mid \text{do}(X))$ non determina $P(Y_{\{x'\}} \mid X, Y)$ senza ipotesi ulteriori sulla struttura del modello. Questa irriducibilità non è un problema computazionale — non si risolve con più dati o con modelli più grandi. È una questione di struttura informazionale: per ragionare causalmente occorre informazione che non è presente in nessuna distribuzione di probabilità, per quanto ricca.

1.1.1 La dimostrazione formale dell'irriducibilità: il controesempio canonico

Il modo più diretto di convincersi che l'irriducibilità della gerarchia causale non è una questione di convenzione terminologica consiste nel costruire esplicitamente una coppia di Modelli Causali Strutturali che producono la medesima distribuzione osservativa pur implicando distribuzioni interventistiche radicalmente distinte. La costruzione che segue, nella sua essenzialità formale, è dovuta in sostanza a Pearl (2009, Capitolo 3), qui esposta con sufficiente completezza da rendere ogni passaggio verificabile.

Si considerino due variabili binarie $X, Y \in \{0, 1\}$ e si postuli che la distribuzione congiunta osservativa soddisfi $P(X = 1) = P(Y = 1) = 1/2$ e $P(X = 1, Y = 1) = 1/2$, ovvero X e Y appaiono perfettamente correlate in ogni campione estratto dalla distribuzione. Il primo modello M_1 specifica le equazioni strutturali $X := U_X$ e $Y := X$, con $U_X \sim \text{Bernoulli}(1/2)$ indipendente. Il grafo causale associato è l'arco singolo $X \rightarrow Y$. Il secondo modello M_2 specifica invece $X := U$ e $Y := U$, con $U \sim \text{Bernoulli}(1/2)$ causa comune latente non osservata. Il grafo causale associato è la biforcazione $X \leftarrow U \rightarrow Y$, priva di qualsiasi freccia diretta tra X e Y .

Entrambi i modelli implicano la stessa distribuzione osservativa: $P(X = 1) = P(Y = 1) = 1/2$ e $P(X = Y) = 1$ con probabilità uno. Nessun test statistico condotto su dati osservativi può distinguere M_1 da M_2 : le due distribuzioni congiunte $P_1(X, Y)$ e $P_2(X, Y)$ sono identiche. Eppure le distribuzioni interventistiche divergono in modo drastico. In M_1 , applicando $\text{do}(X = 0)$ — cioè rimuovendo il meccanismo naturale che determina X e imponendo $X = 0$ per via esogena — l'equazione strutturale $Y := X$ implica immediatamente $Y = 0$. Dunque:

$$P_{\{M_1\}}(Y = 1 \mid \text{do}(X = 0)) = 0$$

In M_2 , l'intervento $\text{do}(X = 0)$ recide l'unico percorso attraverso cui X dipende da U , ma non altera il meccanismo $Y := U$ che rimane intatto. La distribuzione di Y nell'ipotetico mondo in cui X è stato forzato a zero è perciò quella marginale di U , ossia:

$$P_{\{M_2\}}(Y = 1 \mid \text{do}(X = 0)) = P(U = 1) = 1/2$$

La separazione è netta: 0 contro 1/2. Poiché i due modelli sono osservativamente indistinguibili ma interventisticamente opposti, nessuna quantità derivabile dalla sola $P(X, Y)$ — per quanto ricca, per quanto stimata su campioni infiniti — può determinare quale dei due modelli sia corretto. Il divario non è quantitativo ma strutturale: richiede conoscenza dell'architettura causale, non della distribuzione.

1.1.2 L'irriducibilità del controfattuale rispetto all'intervento

Un argomento analogo dimostra che il livello controfattuale non si riduce a quello interventistico. Si costruiscano due modelli che producono distribuzioni interventistiche identiche per ogni valore di X , ma che divergono sulle distribuzioni controfattuali.

Si definisca M_A con le equazioni strutturali $X := U_X$, $Y := f(X, U_Y)$ dove $U_X \sim \text{Bernoulli}(1/2)$, $U_Y \sim \text{Bernoulli}(1/2)$ indipendenti, e $f(x, u) = 1$ se $x = u$, 0 altrimenti. Si verifica che $P_{\{M_A\}}(Y = 1 \mid \text{do}(X = x)) = P(f(x, U_Y) = 1) = P(U_Y = x) = 1/2$ per $x \in \{0, 1\}$. Si definisca M_B con le equazioni $X := U_X$, $Y := 1/2$ con probabilità 1 (Y costante e indipendente da tutto). Anche qui $P_{\{M_B\}}(Y = 1 \mid \text{do}(X = x)) = 1/2$. Le due distribuzioni interventistiche coincidono perfettamente.

Le distribuzioni controfattuali divergono però in modo netto. Si consideri la quantità $P(Y_{\{X=0\}} = 1 \mid X = 1, Y = 1)$: la probabilità che Y sarebbe stato 1 nel mondo controfattuale $X = 0$, dato che nel mondo effettivo si sono osservati $X = 1$ e $Y = 1$. In M_A , l'evento $Y = 1$ con $X = 1$ implica $U_Y = 1$. Nel mondo controfattuale in cui X sarebbe stato 0, l'equazione strutturale darebbe $Y = f(0, U_Y) = f(0, 1) = 0$. Dunque:

$$P_{\{M_A\}}(Y_{\{X=0\}} = 1 \mid X = 1, Y = 1) = 0$$

In M_B invece Y è costantemente $1/2$ indipendentemente da qualsiasi intervento o condizionamento, dunque il controfattuale è identico all'incondizionato:

$$P_{\{M_B\}}(Y_{\{X=0\}} = 1 \mid X = 1, Y = 1) = 1/2$$

La separazione 0 contro $1/2$ si ripete, questa volta al terzo livello della gerarchia contro il secondo. I due modelli condividono l'intera struttura interventistica eppure divergono sui controfattuali: la conoscenza di $P(Y \mid \text{do}(X))$ per ogni X non è sufficiente a determinare $P(Y_{\{X=0\}} \mid X = 1, Y = 1)$ senza ulteriori ipotesi sulla distribuzione congiunta del rumore esogeno — ipotesi che appartengono alla struttura del modello, non alla distribuzione osservativa o interventistica.

1.1.3 Conseguenze per i sistemi JEPA: l'associazione come limite strutturale

Le due dimostrazioni precedenti hanno conseguenze dirette per la valutazione epistemica di qualsiasi sistema di apprendimento, inclusi i sistemi JEPA. Un encoder JEPA addestrato con l'obiettivo di predire rappresentazioni latenti di regioni target a partire da rappresentazioni latenti di regioni contestuali opera, nella sua struttura fondamentale, come un approssimatore della distribuzione

condizionale $P(z_t \mid z_c)$ nello spazio latente. Questa distribuzione appartiene al primo livello della gerarchia causale: è una quantità puramente osservativa, per quanto sofisticata sia la sua parametrizzazione.

Il sistema non può, sulla sola base di questo obiettivo, apprendere la distribuzione interventistica $P(z_t \mid \text{do}(z_c = v))$ per valori v non osservati durante il training. Né, a fortiori, può apprendere distribuzioni controfattuali della forma $P(z_t^{\{c'\}} \mid z_c, z_t)$ che richiederebbero di ragionare su mondi alternativi in cui il contesto fosse stato diverso. Questa limitazione non dipende dall'architettura specifica — dalla scelta del predittore s_ψ , dalla dimensionalità dello spazio latente, dal meccanismo di stop-gradient — ma dalla struttura logica dell'obiettivo: qualsiasi obiettivo che ottimizza una funzione di perdita su coppie (input, target) campionate dalla stessa distribuzione produce al più un approssimatore del primo livello.

La domanda scientifica precisa — quella che motiva l'intera Parte Prima di questa monografia — è dunque: sotto quali condizioni aggiuntive, con quale struttura architetturale supplementare, e attraverso quale obiettivo di addestramento modificato un sistema JEPA può avvicinarsi al secondo livello della gerarchia? Le risposte che la letteratura ha proposto, da iVAE a CausalVAE al framework CV-JEPA sviluppato nel Capitolo 3, convergono tutte sulla stessa strategia fondamentale: arricchire il modello con variabili ausiliarie che rompano la simmetria del problema osservativo e rendano identificabili, nel senso tecnico del termine, i fattori latenti causalmente rilevanti. Il Capitolo 2 svolge questa strategia in dettaglio.

1.2 I Modelli Causali Strutturali: la grammatica formale dell'intervento

La risposta di Pearl al problema dell'irriducibilità causale è il linguaggio dei *Modelli Causali Strutturali* (Structural Causal Models, SCM). Un SCM è una quadrupla

$$M = \langle U, V, F, P(U) \rangle$$

dove $U = \{U_1, \dots, U_k\}$ è un insieme di variabili *esogene* — quelle i cui valori sono determinati da cause esterne al modello, non modellate esplicitamente — e $V = \{V_1, \dots, V_n\}$ è un insieme di variabili *endogene*, le cui relazioni causali sono il soggetto dell'analisi. L'insieme $F = \{f_1, \dots, f_n\}$ raccoglie le *equazioni strutturali*: per ogni variabile endogena V_i esiste una funzione f_i tale che

$$V_i = f_i(\text{Pa}(V_i), U_i)$$

dove $\text{Pa}(V_i) \subseteq V \setminus \{V_i\}$ è l'insieme dei *genitori causali* di V_i nel grafo associato, e U_i è la componente di rumore esogena specifica per V_i . Il grafo causale $G = (V, E)$ è il grafo aciclico diretto (DAG) che si ottiene aggiungendo un arco da V a V_i ogniquale volta $V \in \text{Pa}(V_i)$. La distribuzione $P(U)$ sul rumore esogeno,

combinata con le equazioni strutturali, determina univocamente la distribuzione osservativa $P(V)$ sulle variabili endogene.

L'operatore di intervento $do(V_i = v_i)$ è definito come la sostituzione dell'equazione strutturale di V_i con la costante v_i , ottenendo il *modello mutilato* $M_{\{V_i = v_i\}}$ in cui tutti gli archi entranti in V_i sono rimossi. La distribuzione interventale $P(V \setminus \{V_i\} \mid do(V_i = v_i))$ è la distribuzione marginale di $V \setminus \{V_i\}$ nel modello mutilato. Questa definizione ha una conseguenza fondamentale: le distribuzioni interventali non sono in generale ricavabili dalla distribuzione osservativa $P(V)$ senza conoscere la struttura del DAG G e la forma delle equazioni strutturali — salvo in condizioni specifiche di identificabilità che il do-calcolo permette di verificare algebricamente.

Il *do-calcolo* è un sistema di tre regole di manipolazione delle probabilità condizionali che, applicato ricorsivamente, permette di determinare se e come una distribuzione interventale $P(Y \mid do(X))$ sia identificabile a partire dalla distribuzione osservativa $P(V)$ dato il grafo G . Siano X, Y, Z, W sottoinsiemi disgiunti di V . Le tre regole sono:

Regola 1 (inserzione e cancellazione di osservazioni): se nel grafo mutilato $G_{\{X\}}$ in cui vengono rimossi tutti gli archi entranti in X vale la d-separazione $(Y \perp\!\!\!\perp Z \mid X, W)$, allora

$$P(Y \mid do(X), Z, W) = P(Y \mid do(X), W)$$

Regola 2 (scambio tra azione e osservazione): se nel grafo $G_{\{XZ\}}$ in cui vengono rimossi gli archi entranti sia in X che in Z vale $(Y \perp\!\!\!\perp Z \mid X, W)$, allora

$$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), Z, W)$$

Regola 3 (cancellazione di azioni): se nel grafo $G_{\{XZ(W)\}}$ vale $(Y \perp\!\!\!\perp Z \mid X, W)$, dove $Z(W)$ indica i nodi di Z che non sono antenati di W nel grafo $G_{\{X\}}$, allora

$$P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), W)$$

La completezza del do-calcolo — dimostrata da Shpitser e Pearl nel 2006 — garantisce che qualsiasi distribuzione interventale identificabile dal grafo G possa essere espressa in termini di quantità osservative attraverso l'applicazione ripetuta di queste tre regole. Questo risultato trasforma il problema dell'identificazione causale da questione aperta in algoritmo eseguibile.

1.2.1 Derivazione completa delle tre regole del do-calcolo

Il do-calcolo è un sistema di riscrittura formale che opera su espressioni probabilistiche contenenti l'operatore $do(\cdot)$. La sua completezza — dimostrata da Shpitser e Pearl nel 2006, e poi affinata da Huang e Valtorta (2006) in modo

indipendente — garantisce che qualsiasi distribuzione interventistica identificabile rispetto a un dato grafo causale G possa essere espressa in termini di quantità osservative mediante l'applicazione iterata delle tre regole. Per derivare le regole nella loro forma più precisa, è necessario introdurre alcuni strumenti grafici preliminari.

Dato un DAG G con insieme di nodi V , e dato $X \subseteq V$, si denota con $G_{-}\{X\}$ il grafo mutilato ottenuto rimuovendo tutti gli archi *entranti* in X (ossia, tagliando tutte le frecce causali che puntano a nodi in X), e con $G_{-}\{X\}$ il grafo mutilato ottenuto rimuovendo tutti gli archi *uscenti* da X . La notazione $G_{-}\{X, Z\}$ combina i due: si rimuovono gli archi entranti in X e quelli uscenti da Z . Si denota con $Z(W)$ l'insieme dei nodi in Z che non sono antenati di W nel grafo $G_{-}\{X\}$, dove gli antenati sono calcolati nel grafo mutilato.

Siano X, Y, Z, W quattro insiemi disgiunti di variabili. La d-separazione nel grafo G , denotata $(Y \perp\!\!\!\perp Z \mid X, W)_G$, afferma che ogni cammino non direzionato tra un nodo di Y e un nodo di Z è bloccato da $X \cup W$ nel grafo G . Con questa notazione, le tre regole assumono la seguente forma esatta.

Regola 1 — Inserzione e cancellazione di osservazioni. Se nel grafo mutilato $G_{-}\{X\}$ vale la d-separazione $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{-}\{X\}}$, allora la variabile Z è irrilevante per Y una volta condizionato su (X, W) , nell'ipotetico mondo in cui X è stato forzato per intervento. Formalmente:

$$P(y \mid \text{do}(x), z, w) = P(y \mid \text{do}(x), w)$$

La motivazione intuitiva è che, nel mondo in cui X è stato forzato, tutte le strade causali naturali verso X sono state recise. Le informazioni aggiuntive z quindi non alterano la distribuzione di Y , poiché qualsiasi percorso attraverso cui z potrebbe influenzare y è già bloccato dal condizionamento su w nel grafo mutilato.

Regola 2 — Scambio tra azione e osservazione. Se nel grafo $G_{-}\{X, Z\}$ vale $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{-}\{X, Z\}}$, allora intervenire su Z produce lo stesso effetto che semplicemente osservare Z nel mondo in cui X è già stato forzato. Formalmente:

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), z, w)$$

La condizione di d-separazione richiesta — nel grafo da cui sono stati rimossi sia gli archi entranti in X sia quelli uscenti da Z — garantisce che Z non abbia percorsi backdoor che possano essere confusi con il suo effetto causale diretto su Y . Quando questa condizione è soddisfatta, osservare Z equivale a intervenire su Z : non vi sono confondenti attivi.

Regola 3 — Cancellazione di azioni. Se nel grafo $G_{-}\{X, Z(W)\}$ vale $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{-}\{X, Z(W)\}}$, dove $Z(W) \subseteq Z$ indica i nodi di Z che non sono

antenati di W in $G_{\setminus \{X\}}$, allora l'intervento su Z è irrilevante per Y . Formalmente:

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), w)$$

La Regola 3 cattura la situazione in cui Z non può influenzare Y attraverso nessun cammino che non passi per nodi già bloccati da W . La distinzione tra Z e $Z(W)$ è sottile ma essenziale: alcuni nodi di Z potrebbero essere antenati di W e quindi avere percorsi attivi verso Y attraverso W — questi nodi vengono esclusi da $Z(W)$ perché il loro effetto su Y è già mediato da W e non si azzera semplicemente escludendo l'intervento su di essi.

1.2.2 Il criterio del backdoor e la formula di aggiustamento

Tra le applicazioni più immediate del do-calcolo si trova la *formula di aggiustamento backdoor*, che traduce la distribuzione interventistica $P(y \mid \text{do}(x))$ in un integrale di quantità osservative quando esiste un insieme di covariate Z che soddisfa il criterio del backdoor rispetto alla coppia (X, Y) nel grafo G .

Il *criterio del backdoor* è soddisfatto da un insieme Z rispetto a (X, Y) se e soltanto se valgono due condizioni: (i) nessun nodo in Z è un discendente di X nel grafo G ; (ii) l'insieme Z blocca ogni percorso backdoor tra X e Y nel grafo G . Un percorso backdoor è un cammino non direzionato che comincia con una freccia entrante in X — ossia un cammino della forma $X \leftarrow \cdots \rightarrow Y$ che bypassa il percorso causale diretto $X \rightarrow \cdots \rightarrow Y$ attraverso confondenti.

Quando Z soddisfa il criterio del backdoor, il do-calcolo garantisce che la Regola 2 sia applicabile e che la distribuzione interventistica si possa esprimere come:

$$P(y \mid \text{do}(x)) = \sum_z P(y \mid x, z) \cdot P(z)$$

che nel caso continuo diventa l'integrale:

$$P(y \mid \text{do}(x)) = \int P(y \mid x, z) \cdot P(z) dz$$

Questa formula — nota come *formula di aggiustamento backdoor* o *formula di aggiustamento per confondenti* — traduce un problema causale in un calcolo puramente statistico: richiede soltanto la stima delle distribuzioni condizionali $P(y \mid x, z)$ e della distribuzione marginale $P(z)$ dai dati osservativi, entrambe quantità di primo livello nella gerarchia di Pearl.

La connessione con il paradigma JEPA è diretta. Se lo spazio latente di un encoder JEPA soddisfacesse il criterio del backdoor rispetto alla coppia (azione, rappresentazione futura) — cioè se le componenti del vettore latente z_t soddisfacessero le due condizioni topologiche sopra enumerate rispetto al grafo

causale del processo di generazione dei dati — allora la formula di aggiustamento permetterebbe di computare le distribuzioni interventistiche direttamente dai dati di traiettoria senza mai dover eseguire interventi fisici sull'ambiente. Questa osservazione è il fondamento teorico per cui l'addestramento offline su dati passivi può, in linea di principio, produrre un sistema in grado di pianificare attraverso il do-calcolo latente — un tema che il Capitolo 17 svilupperà in piena formalità.

1.2.1 Derivazione completa delle tre regole del do-calcolo

Il do-calcolo è un sistema di riscrittura formale che opera su espressioni probabilistiche contenenti l'operatore $\text{do}(\cdot)$. La sua completezza — dimostrata da Shpitser e Pearl nel 2006, e poi affinata da Huang e Valtorta (2006) in modo indipendente — garantisce che qualsiasi distribuzione interventistica identificabile rispetto a un dato grafo causale G possa essere espressa in termini di quantità osservative mediante l'applicazione iterata delle tre regole. Per derivare le regole nella loro forma più precisa, è necessario introdurre alcuni strumenti grafici preliminari.

Dato un DAG G con insieme di nodi V , e dato $X \subseteq V$, si denota con $G_{-}\{X\}$ il grafo mutilato ottenuto rimuovendo tutti gli archi *entranti* in X (ossia, tagliando tutte le frecce causali che puntano a nodi in X), e con $G_{-}\{X\}$ il grafo mutilato ottenuto rimuovendo tutti gli archi *uscenti* da X . La notazione $G_{-}\{X, Z\}$ combina i due: si rimuovono gli archi entranti in X e quelli uscenti da Z . Si denota con $Z(W)$ l'insieme dei nodi in Z che non sono antenati di W nel grafo $G_{-}\{X\}$, dove gli antenati sono calcolati nel grafo mutilato.

Siano X, Y, Z, W quattro insiemi disgiunti di variabili. La d-separazione nel grafo G , denotata $(Y \perp\!\!\!\perp Z \mid X, W)_G$, afferma che ogni cammino non direzionato tra un nodo di Y e un nodo di Z è bloccato da $X \cup W$ nel grafo G . Con questa notazione, le tre regole assumono la seguente forma esatta.

Regola 1 — Inserzione e cancellazione di osservazioni. Se nel grafo mutilato $G_{-}\{X\}$ vale la d-separazione $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{-}\{X\}}$, allora la variabile Z è irrilevante per Y una volta condizionato su (X, W) , nell'ipotetico mondo in cui X è stato forzato per intervento. Formalmente:

$$P(y \mid \text{do}(x), z, w) = P(y \mid \text{do}(x), w)$$

La motivazione intuitiva è che, nel mondo in cui X è stato forzato, tutte le strade causali naturali verso X sono state recise. Le informazioni aggiuntive z quindi non alterano la distribuzione di Y , poiché qualsiasi percorso attraverso cui z potrebbe influenzare y è già bloccato dal condizionamento su w nel grafo mutilato.

Regola 2 — Scambio tra azione e osservazione. Se nel grafo $G_{-}\{X, Z\}$ vale $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{-}\{X, Z\}}$, allora intervenire su Z produce lo stesso effetto

che semplicemente osservare Z nel mondo in cui X è già stato forzato. Formalmente:

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), z, w)$$

La condizione di d-separazione richiesta — nel grafo da cui sono stati rimossi sia gli archi entranti in X sia quelli uscenti da Z — garantisce che Z non abbia percorsi backdoor che possano essere confusi con il suo effetto causale diretto su Y . Quando questa condizione è soddisfatta, osservare Z equivale a intervenire su Z : non vi sono confondenti attivi.

Regola 3 — Cancellazione di azioni. Se nel grafo $G_{\setminus \{X, Z(W)\}}$ vale $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\setminus \{X, Z(W)\}}}$, dove $Z(W) \subseteq Z$ indica i nodi di Z che non sono antenati di W in $G_{\setminus \{X\}}$, allora l'intervento su Z è irrilevante per Y . Formalmente:

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), w)$$

La Regola 3 cattura la situazione in cui Z non può influenzare Y attraverso nessun cammino che non passi per nodi già bloccati da W . La distinzione tra Z e $Z(W)$ è sottile ma essenziale: alcuni nodi di Z potrebbero essere antenati di W e quindi avere percorsi attivi verso Y attraverso W — questi nodi vengono esclusi da $Z(W)$ perché il loro effetto su Y è già mediato da W e non si azzerà semplicemente escludendo l'intervento su di essi.

1.2.2 Il criterio del backdoor e la formula di aggiustamento

Tra le applicazioni più immediate del do-calcolo si trova la *formula di aggiustamento backdoor*, che traduce la distribuzione interventistica $P(y \mid \text{do}(x))$ in un integrale di quantità osservative quando esiste un insieme di covariate Z che soddisfa il criterio del backdoor rispetto alla coppia (X, Y) nel grafo G .

Il *criterio del backdoor* è soddisfatto da un insieme Z rispetto a (X, Y) se e soltanto se valgono due condizioni: (i) nessun nodo in Z è un discendente di X nel grafo G ; (ii) l'insieme Z blocca ogni percorso backdoor tra X e Y nel grafo G . Un percorso backdoor è un cammino non direzionato che comincia con una freccia entrante in X — ossia un cammino della forma $X \leftarrow \cdots \rightarrow Y$ che bypassa il percorso causale diretto $X \rightarrow \cdots \rightarrow Y$ attraverso confondenti.

Quando Z soddisfa il criterio del backdoor, il do-calcolo garantisce che la Regola 2 sia applicabile e che la distribuzione interventistica si possa esprimere come:

$$P(y \mid \text{do}(x)) = \sum_z P(y \mid x, z) \cdot P(z)$$

che nel caso continuo diventa l'integrale:

$$P(y \mid \text{do}(x)) = \int P(y \mid x, z) \cdot P(z) dz$$

Questa formula — nota come **formula di aggiustamento backdoor** o **formula di aggiustamento per confondenti** — traduce un problema causale in un calcolo puramente statistico: richiede soltanto la stima delle distribuzioni condizionali $P(y \mid x, z)$ e della distribuzione marginale $P(z)$ dai dati osservativi, entrambe quantità di primo livello nella gerarchia di Pearl.

La connessione con il paradigma JEPA è diretta. Se lo spazio latente di un encoder JEPA soddisfacesse il criterio del backdoor rispetto alla coppia (azione, rappresentazione futura) — cioè se le componenti del vettore latente z_t soddisfacessero le due condizioni topologiche sopra enumerate rispetto al grafo causale del processo di generazione dei dati — allora la formula di aggiustamento permetterebbe di computare le distribuzioni interventistiche direttamente dai dati di traiettoria senza mai dover eseguire interventi fisici sull'ambiente. Questa osservazione è il fondamento teorico per cui l'addestramento offline su dati passivi può, in linea di principio, produrre un sistema in grado di pianificare attraverso il do-calcolo latente — un tema che il Capitolo 17 svilupperà in piena formalità.

1.3 D-separazione e indipendenze condizionali come vincoli sul DAG

La nozione di *d-separazione* è lo strumento chiave per leggere indipendenze condizionali direttamente dal grafo causale senza scorrere l'intera distribuzione congiunta. Dati tre insiemi disgiunti di nodi X, Y, Z in un DAG G , X e Y sono *d-separati* da Z — si scrive $(X \perp\!\!\!\perp Y \mid Z)_G$ — se e soltanto se ogni cammino non direzionato tra un nodo di X e un nodo di Y è *bloccato* da Z . Un cammino è bloccato da Z se contiene almeno una delle seguenti strutture:

una *chain* $A \rightarrow B \rightarrow C$ o una *fork* $A \leftarrow B \rightarrow C$ in cui $B \in Z$ (il nodo centrale è osservato), oppure un *collider* $A \rightarrow B \leftarrow C$ in cui $B \notin Z$ e nessun discendente di B appartiene a Z (il nodo a collisione non è osservato né è antenato di un nodo osservato).

Il *criterio di Markov* — che afferma che ogni variabile endogena in un SCM è condizionalmente indipendente dai propri non-discendenti dato l'insieme dei propri genitori nel DAG — trasforma la struttura del grafo in vincoli verificabili sulla distribuzione congiunta. Simmetricamente, la *faithfulness* — l'ipotesi che le sole indipendenze condizionali nella distribuzione siano quelle implicate dalla d-separazione nel DAG — permette di inferire vincoli sul grafo a partire da indipendenze osservate nei dati. Questi due principi sono il fondamento degli algoritmi di scoperta causale — PC, FCI, GES — che tentano di ricostruire il DAG causale dai soli dati osservativi.

1.3.1 La d-separazione come algoritmo: le strutture di blocco esplicitate

La definizione di d-separazione, pur enunciabile in modo compatto, nasconde una complessità computazionale che merita di essere resa esplicita attraverso la caratterizzazione delle tre strutture elementari che possono bloccare un cammino. Dato un cammino non direzionato π tra due nodi nel DAG G e dato un insieme di condizionamento Z , il cammino è bloccato da Z se e soltanto se contiene almeno uno dei seguenti tre schemi.

Il primo schema è la **catena** (chain): $A \rightarrow B \rightarrow C$ oppure $A \leftarrow B \leftarrow C$, in cui il nodo centrale B appartiene a Z . La spiegazione intuitiva è che B media il flusso di informazione tra A e C : condizionando su B , si recide questa mediazione e A diventa indipendente da C . Formalmente, nel DAG $X \rightarrow Y \rightarrow Z$ si ha che $P(X, Z \mid Y) = P(X \mid Y) \cdot P(Z \mid Y)$, da cui $X \perp\!\!\!\perp Z \mid Y$ segue immediatamente dalla fattorizzazione della congiunta.

Il secondo schema è la **biforcazione** (fork): $A \leftarrow B \rightarrow C$, in cui B appartiene a Z . Anche qui B è mediatore dell'associazione tra A e C : condizionando su B , si rimuove la causa comune e A e C diventano indipendenti. Nel DAG $Y \leftarrow X \rightarrow Z$ la fattorizzazione implica $P(Y, Z \mid X) = P(Y \mid X) \cdot P(Z \mid X)$, da cui $Y \perp\!\!\!\perp Z \mid X$.

Il terzo schema, concettualmente più sottile, è il **collider**: $A \rightarrow B \leftarrow C$, in cui B **non** appartiene a Z e **nessun discendente** di B appartiene a Z . Un collider blocca il cammino quando non è condizionato: in assenza di condizionamento su B , A e C sono indipendenti attraverso B perché B non è una causa comune ma un **effetto** comune. Tuttavia, condizionando su B o su un suo discendente, si **attiva** il cammino attraverso il collider: il fenomeno noto come **Berkson's paradox** o **collider bias**. Formalmente, nel DAG $Y \rightarrow X \leftarrow Z$ si ha $Y \perp\!\!\!\perp Z$ ma non $Y \perp\!\!\!\perp Z \mid X$.

Per i sistemi JEPA, la struttura del collider ha una conseguenza pratica importante che raramente viene discussa nella letteratura. Se l'architettura include un meccanismo di attenzione che produce uno stato aggregato z_AGG come funzione di più rappresentazioni di input $\{z_i\}$, allora z_AGG è un collider nel grafo causale delle rappresentazioni: tutte le viste $\{z_i\}$ convergono verso z_AGG attraverso frecce entranti. Condizionare su z_AGG durante il probing — come avviene implicitamente quando si addestra un classificatore lineare sull'aggregato — attiva i percorsi attraverso z_AGG e può creare dipendenze statistiche spurie tra le rappresentazioni individuali $\{z_i\}$ che non riflettono alcuna relazione causale tra le proprietà dell'input corrispondenti. Questo è il motivo per cui i risultati del probing su z_AGG devono essere interpretati con cautela quando si intende fare affermazioni sulla struttura causale delle singole viste.

1.3.2 Faithfulness, violazioni parametriche e il bias di selezione nei sistemi JEPA

L'assunzione di faithfulness — che le sole indipendenze condizionali nella distribuzione osservativa siano quelle implicate dalla struttura del DAG — è necessaria per la validità degli algoritmi di scoperta causale, ma è sistematicamente violata in alcune configurazioni che emergono naturalmente nel contesto dei sistemi JEPA.

La violazione più rilevante in pratica riguarda l'*invarianza all'azione* come forma di cancellazione parametrica. Un encoder JEPA addestrato a predire la rappresentazione della vista target a partire dalla rappresentazione della vista contestuale, condizionato sull'azione a che porta da una vista all'altra, sviluppa rappresentazioni che tendono a essere invarianti rispetto alle variazioni indotte dall'azione. Questa invarianza è esattamente ciò che si vuole: il modello apprende a fattorizzare le rappresentazioni in una componente che dipende dall'azione (equivariante) e una che non dipende dall'azione (invariante). Ma questa fattorizzazione può produrre indipendenze condizionali nelle rappresentazioni latenti che non corrispondono ad alcuna d-separazione nel DAG causale dei fattori fisici: la cancellazione dei contributi azionali equivarianti da parte dell'aggregatore può far sembrare indipendenti fattori che nel DAG fisico sono connessi attraverso l'azione.

La conseguenza metodologica è che i test di indipendenza condizionale condotti sulle rappresentazioni z_AGG di un sistema JEPA non possono essere utilizzati direttamente per la scoperta della struttura causale del processo fisico che ha generato i dati: riflettono piuttosto la struttura causale che l'architettura ha appreso a *ignorare* ai fini della predizione, non necessariamente quella che il processo fisico implementa. Per recuperare la struttura causale del processo fisico, è necessario condurre i test di indipendenza sulle rappresentazioni equivarianti z_i prima dell'aggregazione — dove la struttura causale azionale è ancora esplicitamente rappresentata — e poi separare la componente azionale dalla componente invariante attraverso procedure di probing causale che utilizzano gli interventi sull'azione come strumento di identificazione.

Per il Causal Representation Learning, il problema è più difficile: non si osservano direttamente le variabili endogene V del DAG, ma soltanto le loro immagini osservative $x = g(V)$ attraverso una funzione di mistura g non invertibile in generale. Il compito è inferire sia il DAG G che la funzione di mistura g dai soli dati x . Questa difficoltà è il cuore del problema dell'apprendimento causale delle rappresentazioni, e occuperà il Capitolo 2 nella sua interezza.

1.4 Il problema dell'identificabilità nelle rappresentazioni latenti

Prima di procedere alla rassegna dei modelli di apprendimento causale delle rappresentazioni, è necessario fissare con precisione il problema dell'identificabilità che queste architetture devono risolvere. Un modello generativo latente è una coppia $(p_{\theta}(x | z), p(z))$ dove $x \in \mathbb{R}^d$ è la variabile osservata, $z \in \mathbb{R}^n$ (tipicamente $n \ll d$) è la variabile latente, e θ parametrizza il decoder $p_{\theta}(x | z)$. Dato un dataset di osservazioni $\{x^{(i)}\}_i$, si addestra il modello massimizzando la verosimiglianza marginale

$$\log p_{\theta}(x) = \log \int p_{\theta}(x | z) p(z) dz$$

o un suo lower bound variazionale (l'ELBO). Il problema dell'identificabilità chiede: sotto quali condizioni la distribuzione $p_{\theta}(x | z) p(z)$ determinata dall'addestramento rispecchia fedelmente la struttura causale sottostante al processo generativo dei dati?

La risposta generale è scoraggiante: senza vincoli aggiuntivi, i parametri $(\theta, p(z))$ di un VAE standard non sono identificabili. Hyvarinen e Pajunen (1999) hanno dimostrato che per modelli di mistura non lineare — il caso che include i VAE — l'indipendenza statistica dei fattori latenti non è sufficiente a garantirne il recupero. In termini più precisi: dati dati da un processo generativo $z \rightarrow x$, esistono infinite parametrizzazioni alternative $(\theta', p'(z'))$ che producono la stessa distribuzione marginale $p(x)$ pur avendo strutture latenti radicalmente diverse. Questo significa che un VAE addestrato a convergenza può avere appreso fattori latenti che non corrispondono a nulla di interpretabile nella struttura causale del mondo.

La letteratura del Causal Representation Learning risponde a questa diagnosi imponendo vincoli aggiuntivi — sull'architettura del decoder, sulla struttura del prior, o sull'utilizzo di variabili ausiliarie — che rompono la simmetria che rende il problema non identificabile nel caso generale. I due approcci più influenti — l'iVAE di Khemakhem et al. e il CausalVAE di Yang et al. — rappresentano due strategie distinte per imporre questi vincoli, e il loro confronto illumina quale tipo di struttura causale ciascuno è in grado di recuperare.

1.4.1 Il teorema di non-identificabilità di Hyvarinen e Pajunen: forma esplicita

Il risultato fondamentale di Hyvarinen e Pajunen (1999) — che esclude l'identificabilità dei fattori latenti di un modello di mistura non lineare senza restrizioni aggiuntive — può essere enunciato nella forma seguente. Sia il processo generativo definito da $x = g(z)$ con $z \sim \prod_i p_i(z_i)$ (fattori indipendenti) e $g : \mathbb{R}^n \rightarrow \mathbb{R}^d$ una funzione invertibile. Si afferma che esiste una famiglia continua di trasformazioni $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ tali che il modello alternativo $x = g(\varphi(z'))$, con $z' \sim p'(z')$ opportunamente scelto, produce la stessa distribuzione marginale $p(x)$ ma con una struttura latente radicalmente diversa.

La costruzione esplicita per il caso gaussiano è la seguente. Si ponga $\varphi(z) = R \cdot z$ dove R è una qualsiasi matrice ortogonale. Se $p(z) = \mathcal{N}(0, I)$, allora il modello alternativo con decoder $g' = g \circ R$ e prior $p'(z') = \mathcal{N}(0, I)$ produce la stessa distribuzione $p(x) = \int p_{\{g\}}(x|z) \cdot \mathcal{N}(z; 0, I) dz$ poiché la distribuzione gaussiana isotropa è invariante rispetto alle rotazioni: $\mathcal{N}(Rz'; 0, I) = \mathcal{N}(z'; 0, I)$ per qualsiasi R ortogonale. La struttura latente del modello alternativo ha i fattori $z'_i = \sum R_{ij} z_j$ — combinazioni lineari dei fattori originali — che sono statisticamente indipendenti (poiché combinazioni lineari ortogonali di gaussiane indipendenti sono ancora gaussiane indipendenti) ma corrispondono a quantità fisiche diverse da z .

Il caso non gaussiano è più complesso. Hyvarinen e Pajunen dimostrano che se le densità marginali $p_i(z_i)$ sono non gaussiane e analitiche, allora l'insieme delle trasformazioni che preservano la distribuzione $P(x)$ si restringe alle sole trasformazioni *elementwise* (funzioni che agiscono separatamente su ciascuna coordinata) e alle permutazioni degli indici. Questo risultato — noto come identificabilità ICA (Independent Component Analysis) nel caso lineare — non si estende al caso non lineare: per decoder non lineari g , esistono trasformazioni non elementwise che preservano $p(x)$ anche con prior non gaussiani. Il teorema di Locatello et al. (2019) formalizza questa impossibilità per i VAE: senza supervisione aggiuntiva o restrizioni architetturali, il training di un VAE non garantisce il recupero dei fattori causali veri nemmeno a meno di trasformazioni elementwise.

L'implicazione per i sistemi JEPA è che il grado di identificabilità ottenibile dipende criticamente dalla struttura delle assunzioni aggiuntive che si è disposti a imporre. Le due monografie precedenti hanno documentato che gli encoder JEPA addestrati con l'obiettivo standard sviluppano rappresentazioni con proprietà di invarianza e generalizzazione notevoli, ma queste proprietà non equivalgono automaticamente all'identificabilità causale nel senso formale. La Parte Prima di questa monografia è dedicata a mostrare quale struttura aggiuntiva — architetturale o di training — è necessaria per chiudere il divario tra le buone proprietà empiriche degli spazi JEPA e le garanzie formali di identificabilità causale che la letteratura di riferimento richiede.

Capitolo 2 — Identificabilità delle rappresentazioni causali: iVAE, CausalVAE e la letteratura correlata

2.1 iVAE: la variabile ausiliaria come leva di identificabilità

L'Identifiable VAE (iVAE) di Khemakhem, Kingma, Monti e Hyvarinen (2020) risponde al problema dell'identificabilità attraverso un'idea concettualmente elegante: arricchire il modello generativo con una variabile ausiliaria osservata u

che modifica il prior sulle variabili latenti, rompendo la simmetria che rende il problema irrisolvibile nel caso unconditional.

Il modello generativo di iVAE è

$$p_{\{\theta, T, \lambda\}}(x, z | u) = p_{\theta}(x | z) \cdot p_{\{T, \lambda\}}(z | u)$$

dove $p_{\{T, \lambda\}}(z | u)$ appartiene alla famiglia esponenziale con statistiche sufficienti $T(z) = (T_1(z_1), \dots, T(z))$ e parametri naturali $\lambda(u) = (\lambda_1(u), \dots, \lambda(u))$ che dipendono dalla variabile ausiliaria:

$$p_{\{T, \lambda\}}(z | u) = \prod_i [h_i(z_i) \cdot \exp(\langle T_i(z_i), \lambda_i(u) \rangle - \log Z_i(u))]$$

La variabile ausiliaria u può essere interpretata come contesto, come dominio, come segmento temporale, o come qualsiasi altra fonte di variazione che modifica sistematicamente la distribuzione latente. La scelta cruciale è che u è osservata durante il training, ma non è necessariamente disponibile durante l'inferenza su nuovi dati: serve durante l'addestramento per rompere la simmetria, non è richiesta per il deployment.

Il training di iVAE ottimizza il lower bound variazionale

$$\mathcal{L}_{\text{iVAE}}(\theta, T, \lambda, \varphi) = \mathbb{E}_{\{q_{\varphi}(z|x, u)\}} [\log p_{\theta}(x | z)] - D_{\text{KL}}(q_{\varphi}(z|x, u) \parallel p_{\{T, \lambda\}}(z|u))$$

dove $q_{\varphi}(z|x, u)$ è l'encoder variazionale che dipende sia dall'osservazione x che dalla variabile ausiliaria u . La presenza di u nell'encoder è necessaria per permettere all'encoder di adattare la propria approssimazione posteriore al prior condizionato.

Il teorema di identificabilità di Khemakhem et al. — il risultato centrale del paper — afferma che, sotto condizioni di regolarità e assumendo che le statistiche sufficienti $T(z)$ formino un sistema *sufficientemente ricco* (tecnicamente: che esista una matrice di coefficienti $L = (\lambda(u^1) - \lambda(u^0), \dots, \lambda(u^n) - \lambda(u^0))$ per $n+1$ punti del supporto di U che sia invertibile), i parametri del modello (θ, T, λ) sono identificabili nel senso seguente: se due modelli (M, M') con parametri (θ, T, λ) e (θ', T', λ') producono la stessa distribuzione congiunta $p(x, u)$ per ogni u nel supporto di U , allora i loro spazi latenti sono collegati da una trasformazione dell'ordine: esistono una permutazione σ di $\{1, \dots, n\}$ e funzioni invertibili $\psi_i : \mathbb{R} \rightarrow \mathbb{R}$ tali che

$$T'\{\sigma(i)\}(\tilde{z}')\{\sigma(i)\} = \psi_i(T_i(z_i)) \text{ per ogni } i$$

In altre parole: le componenti latenti non sono identificabili in modo assoluto, ma sono identificabili *a meno di permutazione e trasformazioni entrywise*, che è il risultato più forte raggiungibile per un modello non lineare di questo tipo. L'identificabilità entrywise garantisce in particolare che se i fattori latenti veri sono causalmente indipendenti tra loro (condizionatamente a u), il modello

addestrato recupererà fattori che rispecchiano questa struttura, a meno del riordinamento delle componenti.

La connessione con JEPA diventa visibile quando si osserva che la variabile ausiliaria u di iVAE ha la stessa struttura funzionale del *contesto* x_c di JEPA: è l'informazione che permette di condizionare il prior sulla variabile bersaglio (z in iVAE, z_t in JEPA). L'encoder di iVAE $q_\varphi(z|x, u)$ è strutturalmente analogo all'encoder target di Var-JEPA che produce z_t condizionato sia all'osservazione x_t che, indirettamente, al contesto z_c attraverso il predittore s_ψ . La differenza fondamentale è che iVAE ottimizza l'identificabilità del modello latente, mentre Var-JEPA ottimizza la predittività della rappresentazione: si tratta di due obiettivi compatibili ma non identici, e la Parte Prima di questa monografia è dedicata a mostrare come possano essere integrati in un framework unico.

2.1.1 Dimostrazione del teorema di identificabilità di iVAE

Il teorema di identificabilità di Khemakhem, Kingma, Monti e Hyvarinen (2020) è il risultato analitico centrale su cui poggia l'intera proposta di iVAE. La sua dimostrazione procede in tre stadi distinti e consecutivi: si stabilisce prima che la verosimiglianza marginale $p(x | u)$ determina univocamente la distribuzione congiunta; si mostra poi che la distribuzione congiunta determina i parametri del modello a meno di certe ambiguità; infine si caratterizza la forma esatta di tali ambiguità come le sole ammissibili dal problema.

Primo stadio. Il modello iVAE specifica $p_{\{\theta, T, \lambda\}}(x, z | u) = p_\theta(x | z) \cdot p_{\{T, \lambda\}}(z | u)$, con $p_{\{T, \lambda\}}(z | u) = \prod_i h_i(z_i) \cdot \exp(T_i(z_i) \cdot \lambda_i(u) - \log Z_i(u))$. Si supponga che due parametrizzazioni (θ, T, λ) e (θ', T', λ') producano la stessa distribuzione $p(x | u)$ per ogni u nel supporto di U . Integrando sulla distribuzione del rumore variazionale e usando l'iniettività del decoder $p_\theta(x | z)$ — l'assunzione che punti distinti dello spazio latente producano distribuzioni distinte sull'osservazione — si deduce l'esistenza di una funzione invertibile $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$ tale che $z' = h(z)$ quasi ovunque rispetto alla distribuzione posteriore.

Secondo stadio. La funzione h deve soddisfare simultaneamente due condizioni derivate dalla fattorizzazione del modello. La prima — compatibilità del decoder: $p_{\{\theta'\}}(x | h(z)) = p_\theta(x | z)$ — impone che il decoder trasformato produca le stesse osservazioni. La seconda — compatibilità del prior: $p_{\{T', \lambda'\}}(h(z) | u) \cdot |\det \partial h / \partial z| = p_{\{T, \lambda\}}(z | u)$ — impone che il prior trasformato sia coerente con il prior originale. Esplicitando la seconda condizione nella forma logaritmica per la famiglia esponenziale:

$$\sum_i T'_i(h_i(z_i)) \cdot \lambda'_i(u) + \log |\det \partial h / \partial z| = \sum_i T_i(z_i) \cdot \lambda_i(u) + \text{costante in } u$$

Terzo stadio. Questa equazione deve valere per ogni u nel supporto di U . Si valuti l'equazione in $n+1$ punti $u^0, u^1, \dots, u^n$ e si sottraggano le valutazioni in u^j e u^0 per $j = 1, \dots, n$, ottenendo il sistema:

$$\sum_i T_i(h_i(z_i)) \cdot (\lambda_i(u^j) - \lambda_i(u^0)) = \sum_i T_i(z_i) \cdot (\lambda_i(u^j) - \lambda_i(u^0)) \quad \forall j=1, \dots, n$$

In forma matriciale, indicando con $L = (\lambda(u^1) - \lambda(u^0) \mid \dots \mid \lambda(u^n) - \lambda(u^0)) \in \mathbb{R}^{n \times n}$ la matrice delle differenze dei parametri naturali, e con $T(z)$ e $T(h(z))$ i vettori delle statistiche sufficienti valutate in z e $h(z)$ rispettivamente, il sistema diventa $L^T \cdot T(h(z)) = L^T \cdot T(z)$. Se L è invertibile — la condizione di sufficienza del teorema — allora $T(h(z)) = T(z)$ per ogni z , ossia le statistiche sufficienti trasformate coincidono con quelle originali. Per la struttura della famiglia esponenziale univariata, questo implica che ciascuna funzione h_i deve essere monotona — una funzione invertibile di $\mathbb{R}$ in $\mathbb{R}$ — e che gli indici possono essere permutati. In simboli: $T_{\sigma(i)}(h_{\sigma(i)}(z_{\sigma(i)})) = \psi_i(T_i(z_i))$ per permutazione σ e funzioni ψ_i invertibili. Questo conclude la dimostrazione: iVAE è identificabile a meno di permutazione degli indici e trasformazioni elementwise monotone delle statistiche sufficienti.

2.2 CausalVAE: equazioni strutturali nello spazio latente

Se iVAE risponde al problema dell'identificabilità sfruttando variabili ausiliarie osservate, CausalVAE di Yang, Liu, Chen e Hu (2021) risponde imponendo una struttura causale esplicita sulle variabili latenti stesse. L'idea è di non limitarsi a richiedere che i fattori latenti siano statisticamente indipendenti — il criterio di disentanglement standard — ma di imporre che siano connessi attraverso un sistema di equazioni strutturali che rispecchia il DAG causale sottostante al processo di generazione dei dati.

Il modello generativo di CausalVAE si articola in due livelli. Il primo è il livello dei fattori causali latenti: si postulano n fattori $\epsilon_1, \dots, \epsilon_n$ indipendenti (le variabili esogene), e n fattori causali $z_1, \dots, z_n$ che sono determinati da equazioni strutturali

$$z_i = \sum_{j \in \text{Pa}(i)} a_{ij} \cdot z_j + \epsilon_i, \text{ per ogni } i \in \{1, \dots, n\}$$

dove $\text{Pa}(i) \subseteq \{1, \dots, n\} \setminus \{i\}$ è l'insieme dei genitori causali di z_i nel DAG, e i coefficienti a_{ij} sono i pesi del DAG. La struttura causale è codificata nella matrice delle adiacenze $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, con la convenzione che $a_{ij} \neq 0$ se e soltanto se $j \rightarrow i$ è un arco del DAG. Il secondo livello è quello osservativo: l'osservazione x è generata da z attraverso un decoder $p_\theta(x \mid z)$.

In forma matriciale, le equazioni strutturali diventano

$$z = A^T \cdot z + \epsilon \Rightarrow z = (I - A^T)^{-1} \epsilon$$

dove la seconda uguaglianza richiede che $(I - A^T)$ sia invertibile, il che è garantito dall'aciclicità del DAG: se G è aciclico, la matrice $(I - A^T)$ è

invertibile con inversa $(I - A^\top)^{-1} = I + A^\top + (A^\top)^2 + \dots$ (la serie di Neumann tronca grazie all'aciclicità).

Il training di CausalVAE ottimizza una funzione di perdita che comprende tre termini:

$$\mathcal{L}_{\text{CausalVAE}} = -\mathbb{E}_{\{q_\varphi(z|x)\}} [\log p_\theta(x|z)] + \beta \cdot D_{\text{KL}}(q_\varphi(z|x) \parallel p_A(z)) + \gamma \cdot h(A)$$

dove il primo termine è il termine di ricostruzione, il secondo è la regolarizzazione KL rispetto al prior causale $p_A(z)$ che tiene conto della struttura del DAG ($p_A(z) = \prod_i p(\epsilon_i)$ dove $\epsilon_i = z_i - \sum_{j \in \text{Pa}(i)} a_j \square z$), e il terzo termine $h(A)$ è il *vincolo di aciclicità* introdotto da Zheng et al. (2018) nell'articolo NOTEARS:

$$h(A) = \text{tr}(e^{\{A \circ A\}}) - n$$

dove $A \circ A$ è il prodotto di Hadamard di A con se stesso (quadrato entrywise), e l'esponenziale è la funzione matriciale. Si verifica che $h(A) = 0$ se e soltanto se A corrisponde a un DAG, e $h(A) > 0$ altrimenti: questo termine viene quindi minimizzato durante il training, guidando A verso una struttura aciclica. Il parametro γ cresce nel corso del training attraverso un aggiornamento di tipo Lagrangiano aumentato, garantendo che il vincolo sia soddisfatto a convergenza.

Un dettaglio architetturale di rilievo: nell'implementazione di Yang et al., l'encoder $q_\varphi(z|x)$ è un encoder standard che produce una rappresentazione $\tilde{z}$ dello spazio osservativo, che viene poi trasformata da un layer causale — una rete neurale che implementa la struttura del DAG — per ottenere la rappresentazione causale z . Questo design permette al modello di separare la domanda «come codificare x in una rappresentazione intermedia» dalla domanda «come strutturare causalmente quella rappresentazione».

2.2.1 La derivazione del vincolo di aciclicità NOTEARS e la sua proprietà di liscenza

Il vincolo di aciclicità $h(A) = \text{tr}(e^{\{A \circ A\}}) - n$ introdotto da Zheng, Aragam, Ravikumar e Xing (2018) nell'articolo NOTEARS è una delle innovazioni metodologiche più significative nel dominio della scoperta causale da dati osservativi degli ultimi anni. La sua proprietà fondamentale — essere uguale a zero se e soltanto se A corrisponde a un DAG, e strettamente positivo altrimenti — lo rende utilizzabile come termine di penalità in un problema di ottimizzazione continua, aggirando il precedente approccio combinatorio che rendeva la ricerca di strutture DAG computazionalmente intrattabile per grafi con più di una decina di nodi.

La dimostrazione che $h(A) = 0 \Leftrightarrow A$ è un DAG utilizza la proprietà della funzione matriciale esponenziale. Per una matrice $A \in \mathbb{R}^{n \times n}$, l'esponenziale

matriciale è definito come $e^A = \sum_{k=0}^{\infty} A^k/k!$, che converge assolutamente per qualsiasi matrice finita. Si ha:

$$\text{tr}(e^{A \circ A}) = \text{tr}(\sum_{k=0}^{\infty} (A \circ A)^k/k!) = n + \text{tr}(A \circ A) + \text{tr}((A \circ A)^2)/2! + \dots$$

dove $A \circ A$ è la matrice con elementi $(A \circ A)_{ij} = A_{ij}^2$. Poiché tutti gli elementi di $A \circ A$ sono non negativi (sono quadrati), anche tutti gli elementi di $(A \circ A)^k$ lo sono per ogni $k \geq 0$. La traccia di una matrice con elementi non negativi è zero se e soltanto se tutti gli elementi diagonali sono zero, ossia se e soltanto se $(A \circ A)^k_{ii} = 0$ per ogni i e ogni $k \geq 1$.

L'elemento $(M^k)_{ii}$ di un grafo di adiacenza $M = A \circ A$ conta il numero di cammini di lunghezza k che partono e tornano al nodo i nel grafo con pesi proporzionali a A_{ij}^2 . Tale elemento è zero per ogni $k \geq 1$ se e soltanto se il grafo non contiene cicli diretti — ossia se e soltanto se A è la matrice di adiacenza di un DAG (aciclico per costruzione). Poiché la somma $\text{tr}(e^{A \circ A}) - n = \sum_{k=1}^{\infty} \text{tr}((A \circ A)^k)/k!$ è una somma di termini non negativi, essa è zero se e soltanto se ogni termine è zero, il che equivale all'assenza di cicli.

La proprietà di lisciazza di h — il fatto che $h : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ sia differenziabile ovunque con gradiente calcolabile efficacemente — è ciò che la rende utilizzabile nell'addestramento di reti neurali. Il gradiente rispetto alla matrice A è:

$$\partial h / \partial A = 2A \circ e^{A \circ A^T}$$

dove il prodotto di Hadamard $\circ$ è componentwise. Questo gradiente può essere calcolato in $O(n^3)$ operazioni mediante algoritmi standard per l'esponenziale matriciale (decomposizione di Padé con scaling e squaring), e può essere propagato attraverso la catena di derivazione dell'obiettivo CausalVAE con il meccanismo standard di backpropagation.

Il metodo di Lagrangiano aumentato adottato da Yang et al. per soddisfare il vincolo $h(A) = 0$ consiste nell'ottimizzare iterativamente la funzione lagrangiana aumentata:

$$L_\gamma(A, \theta, \varphi, \alpha) = \mathcal{L}_{\text{CausalVAE}} + \alpha \cdot h(A) + (\gamma/2) \cdot h(A)^2$$

dove α è il moltiplicatore di Lagrange e γ è il parametro di penalità. Gli aggiornamenti di α e γ seguono lo schema:

$$\alpha \leftarrow \alpha + \gamma \cdot h(A)$$

$$\gamma \leftarrow \beta \cdot \gamma \text{ se } h(A) > \varrho \cdot h(A_{\text{prev}}), \text{ altrimenti } \gamma \text{ invariato}$$

con $\beta > 1$ e $\varrho < 1$ iperparametri del metodo. Questo schema garantisce che, a convergenza, il vincolo $h(A) = 0$ sia soddisfatto con precisione arbitraria, producendo una matrice A che corrisponde esattamente a un DAG. La verifica finale dell'aciclicità — poiché ottimizzazione numerica non garantisce

l'uguaglianza esatta — avviene attraverso la sogliatura degli elementi: si pone $A_{\{ij\}} = 0$ se $|A_{\{ij\}}| < \epsilon$ per una soglia ϵ piccola, e si verifica numericamente che la matrice risultante sia aciclica attraverso l'algoritmo di Kahn per il topological sorting.

2.3 DEAR e il framework delle rappresentazioni disentangled causalmente

La letteratura del causal disentanglement comprende, accanto a iVAE e CausalVAE, una serie di approcci che affrontano il problema da angolazioni diverse. Vale la pena soffermarsi brevemente su un terzo filone — i modelli che impongono la struttura causale sull'equazione strutturale del decoder piuttosto che sull'encoder o sullo spazio latente direttamente — perché chiarisce un aspetto del problema che i due modelli precedenti lasciano in ombra.

In questi approcci, il decoder è strutturato come un sistema causale esplicito: ogni fattore osservato è funzione di un sottoinsieme di fattori latenti causalmente ordinati, e la dipendenza tra fattori osservati e latenti rispecchia il grafo causale che si vuole recuperare. Il vantaggio è che la struttura causale viene imposta nel dominio osservativo, dove è più facile fornire supervisione — ad esempio attraverso annotazioni parziali sui fattori latenti, oppure attraverso pair di osservazioni ottenute da interventi noti. Il limite è che questa strategia funziona bene quando il processo causale è sufficientemente semplice da essere specificato nel decoder, ma diventa computazionalmente intrattabile per DAG complessi o ad alta dimensionalità.

La caratteristica comune a tutti questi approcci è che richiedono o supervisione parziale (etichette su alcuni fattori latenti), o un'assunzione strutturale forte sulla forma delle equazioni strutturali (tipicamente linearità o bilinearità), o entrambe. Nessuno di essi è un metodo puramente non supervisionato: c'è sempre un trade-off tra la forza delle assunzioni e la quantità di supervisione richiesta. Questo trade-off è un vincolo fondamentale — non contingente — del Causal Representation Learning, e il suo riconoscimento esplicito è indispensabile per valutare correttamente cosa si può aspettare dall'integrazione con JEPA.

2.3.1 Bound formale di identificabilità nei modelli con struttura causale nel decoder: il caso bilineare e le condizioni minimali di supervisione

La sezione 2.3 ha delineato il terzo filone del Causal Representation Learning — quello che impone la struttura causale nel decoder piuttosto che nell'encoder o nello spazio latente — indicandone il vantaggio principale (supervisione nel dominio osservativo) e il limite principale (intrattabilità per DAG complessi). Si tratta ora di rendere queste affermazioni precise, derivando le condizioni formali di identificabilità per la classe di modelli bilineari, che è quella per cui la teoria è più sviluppata e i risultati più netti.

Sia $z = (z_1, z_2, \dots, z_m) \in \mathbb{R}^m$ il vettore di fattori latenti, con struttura causale descrivibile da un DAG $G = (V, E)$ avente matrice di adiacenza pesata $B \in \mathbb{R}^{m \times m}$ (aciclica per costruzione). Le equazioni strutturali nello spazio latente sono:

$$z_i = \sum_{j \in pa(i)} B_{ij} \cdot z_j + N_i$$

dove N_i è rumore esogeno indipendente con distribuzione $q_i(N_i)$ non-gaussiana (l'ipotesi di non-gaussianità è essenziale per l'identificabilità, come si vedrà). In forma vettoriale: $z = B z + N$, ovvero $(I - B) z = N$, da cui $z = (I - B)^{-1} N =: A N$, con A la matrice degli effetti totali. Il modello del decoder è:

$$x = g(z) + \epsilon$$

dove $g : \mathbb{R}^m \rightarrow \mathbb{R}^d$ è la funzione di generazione osservativa e $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ è il rumore di osservazione. La struttura causale è codificata in B (o equivalentemente in A), non nell'encoder.

Il caso bilineare si ottiene specializzando g alla forma:

$$x = C \cdot (z \otimes u) + \epsilon$$

dove $u \in \mathbb{R}^p$ è una variabile ausiliaria osservata (annotazione parziale, condizione sperimentale, identità del dominio), $\otimes$ è il prodotto tensoriale, e $C \in \mathbb{R}^{d \times (m \cdot p)}$ è la matrice del decoder bilineare. Questa struttura, introdotta da Brehmer et al. (2022) nel contesto del causal world model, generalizza la formulazione lineare consentendo che i fattori latenti interagiscano con la variabile ausiliaria nel produrre l'osservazione — una forma di mixing condizionale che è più realistica del decoder lineare ma computazionalmente trattabile.

Teorema di identificabilità bilineare. Sia $p(x | u)$ la distribuzione delle osservazioni condizionata sulla variabile ausiliaria u , generata dal modello bilineare con fattori latenti z distribuiti secondo $p(z) = \prod_i q_i(z_i)$ (distribuzione a priori fattorizzata con margini non-gaussiani). Sotto le seguenti condizioni:

(B1) La matrice del decoder C ha rango pieno $m \cdot p$: il mapping da $(z \otimes u)$ a x è iniettivo. (B2) La variabile ausiliaria u varia in modo sufficientemente ricco: il rango del tensor $\sum_u p(u) \cdot (u u^T) \otimes I_m$ è pari a $m \cdot p$. (B3) I margini q_i sono non-gaussiani: per ogni i , la funzione caratteristica $\varphi_i(t) = \mathbb{E}[\exp(it z_i)]$ non è mai della forma $\exp(-t^2 \sigma^2 / 2 + i \mu t)$. (B4) Il DAG G è identificabile dalla distribuzione interventale: per ogni intervento $do(z_i = \tilde{z}_i)$, la distribuzione osservativa risultante $p(x | do(z_i = \tilde{z}_i), u)$ è distinguibile da $p(x | u)$ per qualche valore di u .

Allora il modello bilineare è identificabile modulo un gruppo di ambiguità residua $\mathcal{G} = \mathcal{P} \times \mathcal{F}$, dove $\mathcal{P}$ è il gruppo delle permutazioni degli indici latenti e $\mathcal{F}$ è il gruppo delle trasformazioni per-componente invertibili: due soluzioni $(C, B, \{q_{-i}\})$ e $(C', B', \{q'_{-i}\})$ che inducono la stessa distribuzione osservativa $p(x | u)$ sono necessariamente legate da $(C', B', \{q'_{-i}\}) = \pi \circ \varphi \circ (C, B, \{q_{-i}\})$ per qualche permutazione $\pi \in \mathcal{P}$ e sequenza di trasformazioni invertibili $\varphi = (\varphi_{-1}, \dots, \varphi_{-m}) \in \mathcal{F}$.

La dimostrazione procede in tre passi. Nel primo passo si sfrutta la condizione (B1) per invertire il decoder bilineare: per ogni coppia osservata (x, u) con rango pieno di C , il prodotto $(z \otimes u)$ è identificabile da x come $(z \otimes u) = (C^T C)^{-1} C^T x$. Nel secondo passo si usa la struttura tensoriale per isolare z : poiché $z \otimes u$ è il prodotto tensoriale di due vettori e u è osservato, il vettore z è recuperabile dalla decomposizione di rango 1 del tensore osservato condizionato su u — esattamente la condizione (B2) garantisce che questa decomposizione sia unica. Nel terzo passo si usa (B3) e il teorema di Darrois-Skitovic per identificare i fattori latenti: se la distribuzione fattorizzata delle sorgenti indipendenti non-gaussiane è l'unica a produrre la distribuzione osservativa attraverso un mixing lineare invertibile, allora i fattori sono identificati modulo permutazione e riscaldamento. La condizione (B4) aggiunge l'identificabilità del DAG: le distribuzioni interventali distinguono la struttura B dalla matrice A degli effetti totali attraverso il trucco della truncation factor (il DAG mutilato G_{-i} ha matrice di adiacenza B con la colonna i azzerata, e la differenza tra $p(x | u)$ e $p(x | \text{do}(z_{-i} = \tilde{z}_{-i}), u)$ rivela questa struttura). $\square$

Il bound sulla quantità di supervisione necessaria emerge dalla condizione (B2). Si può mostrare che il rango del tensor $\sum_u p(u) \cdot (u u^T) \otimes I_m$ raggiunge $m \cdot p$ se e solo se la variabile ausiliaria u assume almeno $m + 1$ valori distinti in posizione generale — la cosiddetta condizione di "m-span" introdotta da Khemakhem et al. per iVAE e qui estesa al caso bilineare. Questo significa che una supervisione con $m + 1$ condizioni sperimentali distinte (ambienti di raccolta dati, configurazioni del DAG, istanze del dataset) è necessaria e sufficiente per l'identificabilità dei fattori latenti. Il trade-off tra assunzioni strutturali e supervisione richiesta assume qui una forma quantitativa precisa: se m è il numero di fattori latenti che si vuole identificare, la supervisione minima è $m + 1$ condizioni. Ogni assunzione aggiuntiva sulla struttura di B (sparsità, gerarchicità, linearità dei meccanismi) riduce il numero di condizioni necessarie, al limite fino a 2 per modelli bilineari con struttura triangolare nota a priori.

Nel contesto dell'integrazione con JEPa, questo bound ha un'implicazione architettonica diretta. Se il DAG latente di CV-JEPa ha m fattori causali — tipicamente $m \in \{5, 10, 20\}$ nei dataset di riferimento come 3DIEBench — allora un dataset di preaddestramento che copra almeno $m + 1$ ambienti con distribuzioni di input genuinamente diverse soddisfa la condizione di identificabilità bilineare. Il problema empirico è che JEPa non addestra un decoder (per definizione: l'obiettivo JEPa non include mai la ricostruzione dell'osservazione), cosicché la condizione (B1) non è direttamente verificabile. Il

modo di aggirarla è sostituire il decoder bilineare con il predittore JEPA nello spazio latente: l'identificabilità si sposta dall'inverso del decoder all'inverso del predittore, e la condizione di rango pieno (B1) diventa una condizione di biettività del predittore $g_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ tra rappresentazioni di contesto e target — una proprietà che è plausibile nel regime di preaddestramento, ma che richiede verifica formale caso per caso.

Capitolo 3 — Il DAG generativo di Var-JEPA e il do-calcolo latente

3.1 La struttura latente di Var-JEPA come SCM implicito

La dimostrazione centrale di Gögl e Yau è che il JEPA deterministico canonico è il limite di un modello generativo variazionale quando le varianze dei posteriori tendono a zero. Questo risultato, discusso in dettaglio nella seconda monografia del trittico, ha una conseguenza per il Causal Representation Learning che non è stata sfruttata nel paper originale: il modello generativo variazionale di Var-JEPA ha la struttura di un SCM parzialmente specificato, e rendere esplicita questa struttura è il punto di partenza naturale per l'integrazione causale.

Il modello generativo di Var-JEPA può essere scritto come:

$\epsilon_c, \epsilon_t \sim p(\epsilon)$ (rumore esogeno, componenti indipendenti) $z_c = f_c(x_c, \epsilon_c)$ (encoder del contesto con rumore variazionale) $z_t = f_t(x_t, \epsilon_t)$ (encoder del target con rumore variazionale) $\hat{x}_c = g_c(z_c)$ (decoder del contesto) $\hat{x}_t = g_t(z_t)$ (decoder del target) $\hat{z}_t = s_\psi(z_c)$ (predittore: stima di z_t a partire da z_c)

dove f_c, f_t sono le reti encoder variazionali che implementano le distribuzioni posteriori $q_\varphi(z_c|x_c) = \mathcal{N}(\mu_c(x_c), \sigma_c^2(x_c) \cdot I)$ e $q_\varphi(z_t|x_t) = \mathcal{N}(\mu_t(x_t), \sigma_t^2(x_t) \cdot I)$, e s_ψ è la rete predittrice che opera nello spazio latente. Il grafo causale implicito ha i seguenti archi: $\epsilon_c \rightarrow z_c, x_c \rightarrow z_c, z_c \rightarrow \hat{x}_c, \epsilon_t \rightarrow z_t, x_t \rightarrow z_t, z_t \rightarrow \hat{x}_t, z_c \rightarrow \hat{z}_t$.

Si osservi che questo grafo ha una struttura causale specifica: z_c e z_t non sono causalmente connessi nel modello generativo (non c'è arco da z_c a z_t , né viceversa), ma il predittore s_ψ apprende a stimare z_t a partire da z_c sfruttando le dipendenze statistiche che esistono tra le due rappresentazioni nel dato di training. In altri termini: il modello generativo assume che z_c e z_t siano causalmente indipendenti dato il contenuto dell'osservazione, ma apprende a sfruttare la correlazione osservativa tra di essi per la predizione.

Questa struttura è epistemicamente soddisfacente? La risposta dipende dal tipo di relazione che si vuole catturare. Se z_c e z_t sono due viste della stessa scena

fisica, la correlazione tra di essi è prodotta dall'oggetto fisico comune che li ha causati — chiamiamolo s , il *fattore causale condiviso*. Il modello generativo corretto, in questo caso, avrebbe la struttura:

$s \sim p(s)$ (fattore causale latente condiviso) $z_c = f_c(s, \epsilon_c)$ (context encoding come funzione del fattore comune e rumore) $z_t = f_t(s, \epsilon_t)$ (target encoding come funzione del fattore comune e rumore)

In questa struttura, la correlazione tra z_c e z_t è un artefatto della loro dipendenza da s , non una relazione causale diretta tra di essi. Un sistema che vuole *capire* la scena — piuttosto che limitarsi a predirla — dovrebbe inferire s esplicitamente, non soltanto catturare la correlazione tra z_c e z_t . La differenza tra questi due approcci è precisamente la differenza tra apprendimento correlazionale e apprendimento causale.

3.2 Formalizzare il DAG causale sui fattori latenti

Per costruire un'integrazione formale tra Var-JEPA e il Causal Representation Learning, è necessario estendere il grafo causale implicito di Var-JEPA includendo esplicitamente i fattori causali latenti che producono le osservazioni. Si propone il seguente modello generativo esteso, che chiameremo *Causal Var-JEPA* (CV-JEPA):

$s \sim p(s) = \prod_i p(s_i)$ (fattori causali latenti indipendenti, n componenti) $c = A_c \cdot s + \epsilon_c$ (fattori di contesto come trasformazione lineare causale con rumore) $t = A_t \cdot s + \epsilon_t$ (fattori di target come trasformazione lineare causale con rumore) $x_c = g_c(c)$ (osservazione di contesto) $x_t = g_t(t)$ (osservazione di target) $z_c = \text{enc}_c(x_c)$ (rappresentazione di contesto) $z_t = \text{enc}_t(x_t)$ (rappresentazione di target)

dove $A_c, A_t \in \mathbb{R}^{n \times n}$ sono matrici che codificano come i fattori causali contribuiscono ai fattori di ciascuna vista, e $\epsilon_c, \epsilon_t \in \mathbb{R}^n$ sono rumori vista-specifici indipendenti da s . Il DAG del modello generativo ha ora $n + 4$ nodi: $\{s_1, \dots, s_n, c, t, x_c, x_t\}$ più i nodi derivati per le rappresentazioni.

In questo modello, la *correlazione* tra z_c e z_t deriva dalla loro dipendenza comune da s attraverso le trasformazioni A_c e A_t . Il predittore s_ψ di Var-JEPA, nell'ottica causale, è uno stimatore di z_t che sfrutta z_c come variabile proxy di s : si ha $\hat{z}_t = s_\psi(z_c) \approx \mathbb{E}[z_t \mid z_c] = \mathbb{E}[\text{enc}_t(g_t(A_t \cdot s + \epsilon_t)) \mid \text{enc}_c(g_c(A_c \cdot s + \epsilon_c)) = z_c]$.

La domanda causale che CV-JEPA dovrebbe rispondere è: se si interviene su uno specifico fattore causale s_i — forzando $\text{do}(s_i = v)$ — come cambia la distribuzione di z_t ? Questa quantità, che chiameremo *distribuzione predittiva interventale*

$P(z_t \mid \text{do}(s_i = v), z_c)$

non è in generale computabile da Var-JEPA standard, perché Var-JEPA non ha accesso esplicito ai fattori s . Ma se i codificatori enc_c e enc_t sono sufficientemente buoni da far sì che le componenti di z_c e z_t rispecchino i fattori s (nel senso dell'identificabilità iVAE), allora la distribuzione interventale può essere approssimata nel seguente modo: si identifica il sottoinsieme di componenti di z_c che corrisponde a s_i attraverso probing causale (Capitolo 12), si imposta quella componente al valore v , e si propaga attraverso il predittore s_ψ per ottenere $\hat{z}_t(v)$.

3.3 L'ELBO causale: estendere i cinque termini di Var-JEPA

Il training objective di CV-JEPA deve essere una funzione di perdita che: (a) massimizza la verosimiglianza delle osservazioni dato il modello generativo; (b) impone che i fattori latenti z_c e z_t catturino i fattori causali s ; (c) garantisce che il predittore s_ψ stimi correttamente le dipendenze interventali, non soltanto quelle osservative.

Il lower bound variazionale del modello CV-JEPA si ottiene introducendo una distribuzione variazionale $q_\varphi(s, c, t \mid x_c, x_t)$ che approssima il posteriore congiunto. Per fattorizzare il calcolo in modo computazionalmente trattabile, si assume che il posteriore variazionale sia fattorizzato come

$$q_\varphi(s, c, t \mid x_c, x_t) = q_\varphi^s(s \mid z_c, z_t) \cdot q_\varphi^c(c \mid x_c) \cdot q_\varphi^t(t \mid x_t)$$

dove $z_c = enc_c(x_c)$ e $z_t = enc_t(x_t)$ sono le uscite degli encoder. L'ELBO si sviluppa in sette termini:

$$\begin{aligned} \mathcal{L}_{CV-JEPA} = & \mathbb{E}_{\{q_\varphi\}}[\log p_\theta(x_c \mid c)] \text{ [ricostruzione contesto]} + \\ & \mathbb{E}_{\{q_\varphi\}}[\log p_\theta(x_t \mid t)] \text{ [ricostruzione target]} - D_{KL}(q_\varphi^c(c \mid x_c) \parallel p(c \mid s)) \text{ [regolarizzazione contesto]} - D_{KL}(q_\varphi^t(t \mid x_t) \parallel p(t \mid s)) \\ & \text{ [regolarizzazione target]} - D_{KL}(q_\varphi^s(s \mid z_c, z_t) \parallel p(s)) \text{ [regolarizzazione fattori causali]} - \mathbb{E}_{\{q_\varphi\}}[D_{KL}(q_\varphi^t(t \mid x_t) \parallel q_\psi(t \mid z_c))] \text{ [errore predittivo]} \\ & - \gamma \cdot h(A_c, A_t) \text{ [vincolo di struttura causale]} \end{aligned}$$

I primi sei termini sono un'estensione naturale dell'ELBO a cinque termini di Var-JEPA: i termini di ricostruzione e le tre KL corrispondono ai cinque termini originali (con la separazione tra fattori condivisi s e fattori vista-specifici c, t che introduce il quinto termine sulle KL), e il termine predittivo è l'analogo causale della loss JEPA che penalizza la distanza tra la rappresentazione predetta e quella effettiva del target. Il settimo termine $h(A_c, A_t)$ è il vincolo di aciclicità sulle matrici di connessione, analogo al termine NOTEARS di CausalVAE, che garantisce che A_c e A_t definiscano strutture acicliche.

In forma compatta, usando la notazione dell'ELBO surgery di Mathieu et al.:

$$\mathcal{L}_{CV-JEPA} = \mathcal{L}_{recon} + \mathcal{L}_{regol} + \mathcal{L}_{pred} + \mathcal{L}_{causal}$$

dove $\mathcal{L}_{\text{recon}} = \mathbb{E}[\log p_{\theta}(\mathbf{x}_c | \mathbf{c})] + \mathbb{E}[\log p_{\theta}(\mathbf{x}_t | \mathbf{t})]$ è il termine di verosimiglianza, $\mathcal{L}_{\text{regol}} = -D_{\text{KL}}(q \parallel p_s) - D_{\text{KL}}(q \parallel p_c) - D_{\text{KL}}(q \parallel p_t)$ è la somma dei termini di regolarizzazione KL, $\mathcal{L}_{\text{pred}} = -\mathbb{E}[D_{\text{KL}}(q^t \parallel q_{\psi})]$ è il termine predittivo JEPA, e $\mathcal{L}_{\text{causal}} = -\gamma \cdot h(A_c, A_t)$ è il vincolo causale.

Questa derivazione mostra che CV-JEPA non è una proposta *ad hoc* ma emerge naturalmente come l'ELBO del modello generativo CV-JEPA, e che i termini aggiuntivi rispetto a Var-JEPA hanno una giustificazione probabilistica precisa. Il costo computazionale aggiuntivo è concentrato nel calcolo di $h(A_c, A_t)$, che richiede il calcolo dell'esponenziale matriciale di $A \circ A$ — un'operazione di costo $O(n^3)$ dove n è la dimensione dello spazio latente causale, tipicamente molto inferiore alla dimensione dello spazio latente visivo.

3.3.1 Derivazione completa dell'ELBO causale a sette termini

La derivazione dell'ELBO di CV-JEPA dalla definizione del modello generativo procede attraverso l'applicazione standard della disuguaglianza di Jensen alla log-verosimiglianza marginale. La log-verosimiglianza del modello CV-JEPA per una coppia di osservazioni $(\mathbf{x}_c, \mathbf{x}_t)$ è:

$$\log p_{\theta}(\mathbf{x}_c, \mathbf{x}_t) = \log \iiint p_{\theta}(\mathbf{x}_c, \mathbf{x}_t | \mathbf{c}, \mathbf{t}) \cdot p_A(\mathbf{c}, \mathbf{t} | \mathbf{s}) \cdot p(\mathbf{s}) \, ds \, d\mathbf{c} \, d\mathbf{t}$$

dove il modello generativo fattorizza come $p_{\theta}(\mathbf{x}_c | \mathbf{c}) \cdot p_{\theta}(\mathbf{x}_t | \mathbf{t}) \cdot p(\mathbf{c} | \mathbf{s}) \cdot p(\mathbf{t} | \mathbf{s}) \cdot p(\mathbf{s})$. Introducendo la distribuzione variazionale $q_{\psi}(\mathbf{s}, \mathbf{c}, \mathbf{t} | \mathbf{x}_c, \mathbf{x}_t)$ e applicando la disuguaglianza di Jensen attraverso il logaritmo — che è una funzione concava — si ottiene il lower bound variazionale:

$$\log p_{\theta}(\mathbf{x}_c, \mathbf{x}_t) \geq \mathbb{E}_{q_{\psi}} [\log p_{\theta}(\mathbf{x}_c, \mathbf{x}_t, \mathbf{s}, \mathbf{c}, \mathbf{t}) - \log q_{\psi}(\mathbf{s}, \mathbf{c}, \mathbf{t} | \mathbf{x}_c, \mathbf{x}_t)]$$

Espandendo il logaritmo della distribuzione congiunta del modello e della distribuzione variazionale, e usando la fattorizzazione assunta $q_{\psi} = q_{\psi}^s(\mathbf{s} | \mathbf{z}_c, \mathbf{z}_t) \cdot q_{\psi}^c(\mathbf{c} | \mathbf{x}_c) \cdot q_{\psi}^t(\mathbf{t} | \mathbf{x}_t)$, si ottiene l'ELBO nella forma sviluppata:

$$\begin{aligned} \mathcal{L}_{\text{CV-JEPA}} = & \mathbb{E}_{q_{\psi}} [\log p_{\theta}(\mathbf{x}_c | \mathbf{c})] \quad [\text{T}_1: \text{ricostruzione contesto}] \\ & + \mathbb{E}_{q_{\psi}} [\log p_{\theta}(\mathbf{x}_t | \mathbf{t})] \quad [\text{T}_2: \text{ricostruzione target}] \\ & - \mathbb{E}_{q_{\psi}^s} [D_{\text{KL}}(q_{\psi}^c(\mathbf{c} | \mathbf{x}_c) \parallel p(\mathbf{c} | \mathbf{s}))] \quad [\text{T}_3: \text{regolarizzazione c}] \\ & - \mathbb{E}_{q_{\psi}^s} [D_{\text{KL}}(q_{\psi}^t(\mathbf{t} | \mathbf{x}_t) \parallel p(\mathbf{t} | \mathbf{s}))] \quad [\text{T}_4: \text{regolarizzazione t}] \\ & - D_{\text{KL}}(q_{\psi}^s(\mathbf{s} | \mathbf{z}_c, \mathbf{z}_t) \parallel p(\mathbf{s})) \quad [\text{T}_5: \text{regolarizzazione s}] \\ & - \mathbb{E}_{q_{\psi}} [D_{\text{KL}}(q_{\psi}^t(\mathbf{t} | \mathbf{x}_t) \parallel q_{\psi}(\mathbf{t} | \mathbf{z}_c))] \quad [\text{T}_6: \text{predittivo}] \\ & - \gamma \cdot h(A_c, A_t) \quad [\text{T}_7: \text{aciclicità}] \end{aligned}$$

Il termine T_1 e T_2 misurano la qualità della ricostruzione delle osservazioni dai fattori latenti: un modello che comprime bene le osservazioni in rappresentazioni latenti produrrà log-verosimiglianze di ricostruzione alte. I termini T_3 e T_4 regolarizzano i fattori vista-specifici c e t verso la distribuzione implicata dai fattori causali condivisi s — questa regolarizzazione forza i fattori c e t a essere connessi ai fattori causali s attraverso le equazioni strutturali $A_c \cdot s$ e $A_t \cdot s$, e non a essere fattori arbitrari indipendenti. Il termine T_5 regolarizza i fattori causali s verso il prior indipendente $p(s) = \prod_i p(s_i)$, realizzando l'indipendenza causale dei fattori condivisi. Il termine T_6 è il termine predittivo specifico di JEPA: penalizza la distanza tra la distribuzione variazionale sul target $q_\varphi^t(t \mid x_t)$ e la previsione del predittore $q_\psi(t \mid z_c)$, forzando il predittore ad apprendere una rappresentazione delle transizioni causali. Il termine T_7 impone l'aciclicità delle matrici strutturali A_c e A_t attraverso il vincolo NOTEARS.

La derivazione mostra che CV-JEPA non è una proposta *ad hoc* costruita concatenando pezzi di architetture preesistenti: emerge come l'ELBO del modello generativo CV-JEPA, nella stessa identica maniera in cui l'ELBO di Var-JEPA emerge dal modello generativo variazionale di Gögl e Yau, e i sette termini dell'ELBO hanno ciascuno una giustificazione probabilistica precisa che riflette la struttura del modello generativo sottostante.

3.3.2 Condizioni di convergenza e identificabilità in CV-JEPA

Un'analisi dell'ELBO di CV-JEPA permette di derivare le condizioni sotto cui il training converge verso rappresentazioni causalmente identificabili. Si considerino i casi limite dell'ottimizzazione.

Se il termine T_5 domina — ossia se la regolarizzazione sui fattori causali s è molto forte — il modello collasserà verso il prior $p(s)$, producendo fattori s indistinguibili indipendentemente dalle osservazioni. Questo è il collasso rappresentazionale, analogo al collasso del posteriore nei VAE con β molto grande. Per prevenirlo, il peso del termine T_5 deve essere bilanciato con i termini di ricostruzione T_1 e T_2 .

Se il termine T_6 domina — ossia se il peso del termine predittivo è molto grande — il modello forzerà il predittore q_ψ a essere molto vicino alla distribuzione variazionale q_φ^t , indipendentemente dalla qualità della rappresentazione causale. Questo produce un sistema dove il predittore "copia" la distribuzione variazionale piuttosto che apprenderla causalmente dal contesto z_c . Il meccanismo di stop-gradient, già presente nell'architettura originale JEPA, serve esattamente a prevenire questo comportamento degenerare.

In condizioni di ottimizzazione ben bilanciate, la convergenza di CV-JEPA verso una soluzione identificabile richiede che le tre condizioni iVAE (discusse nel Capitolo 2) siano soddisfatte rispetto alle azioni come variabili ausiliarie, e in

aggiunta che il vincolo di aciclicità T_7 converga a zero — condizione garantita dal metodo di Lagrangiano aumentato a patto che l'ottimizzazione raggiunga un punto stazionario del problema aumentato. La verifica empirica di queste condizioni di convergenza attraverso metriche di probing causale (Capitolo 12) è parte integrante del protocollo sperimentale del Sotto-programma I descritto nel Capitolo 20.

Capitolo 4 — Il neuro-simbolismo di terza generazione: NS-CL e l'integrazione con lo spazio latente JEPA

4.1 La gerarchia dell'astrazione concettuale in NS-CL

Il Neuro-Symbolic Concept Learner (NS-CL) di Mao, Gan, Kohli, Tenenbaum e Wu (2019) rappresenta un punto di incontro tra due tradizioni che si sono sviluppate in parallelo senza convergere per decenni: la tradizione del deep learning, che apprende rappresentazioni distribuite a partire da grandi quantità di dati visivi, e la tradizione dell'AI simbolica, che manipola rappresentazioni discrete con struttura logica esplicita. NS-CL non è un'architettura di sintesi nel senso debole — non si limita a combinare moduli preaddestrati — ma propone un framework di addestramento *congiunto* in cui la componente percettiva e la componente simbolica si co-addestrano attraverso una loss che emerge dal rapporto tra domande in linguaggio naturale e risposte su scene visive.

L'architettura di NS-CL si articola su tre livelli funzionali distinti. Il primo è il livello percettivo: una rete neurale convoluzionale (o, nelle versioni più recenti, un vision transformer) estrae rappresentazioni visive degli oggetti presenti nella scena. Il punto cruciale è che queste rappresentazioni sono orientate agli oggetti — l'encoder produce un insieme $\{o_1, \dots, o_n\}$ di vettori di oggetto, non una rappresentazione globale della scena — e i concetti visivi (colore, forma, dimensione, materiale) sono appresi come classificatori lineari nello spazio delle rappresentazioni degli oggetti. Il secondo livello è quello del programma simbolico: ogni domanda in linguaggio naturale viene tradotta in un programma eseguibile strutturato come albero di operazioni simboliche — Filter, Query, Relate, Count, And, Or, Exist — con semantica precisamente definita sulle rappresentazioni visive. Il terzo livello è quello dell'interfaccia neurosimbolica: le operazioni simboliche operano su distribuzioni di probabilità sugli insiemi di oggetti, permettendo all'addestramento end-to-end di propagare gradienti attraverso le operazioni discrete del programma simbolico.

Il vantaggio di NS-CL rispetto a un sistema puramente neurale è la *sistematicità* della rappresentazione: un sistema che ha appreso il concetto di «rosso» e il concetto di «cubo» può rispondere a domande su «cubi rossi» anche se non ha mai visto un cubo rosso durante l'addestramento, perché la composizione dei

concetti è gestita dal programma simbolico. Questa proprietà — che Fodor e Pylyshyn avevano identificato come caratteristica necessaria di qualsiasi sistema cognitivo genuino — è emersa *naturalmente* dall'architettura senza essere codificata esplicitamente nel design.

4.2 L'integrazione strutturale con lo spazio latente JEPA

La connessione tra NS-CL e il framework JEPA non è superficiale. Per comprenderla, occorre identificare il punto di frizione tra i due approcci che li rende complementari piuttosto che alternativi.

NS-CL apprende concetti visivi come classificatori lineari nello spazio delle rappresentazioni degli oggetti. Ma questi classificatori sono addestrati su coppie (scena, domanda-risposta), e dipendono dalla qualità della rappresentazione degli oggetti prodotta dal modulo percettivo. Se quel modulo percettivo è addestrato con un obiettivo puramente supervisionato — predire la risposta alla domanda — le rappresentazioni che esso produce possono essere ottimizzate per il task specifico piuttosto che per la generalità che il reasoning simbolico richiederebbe.

JEPA offre una soluzione naturale: se il modulo percettivo di NS-CL è sostituito da un encoder JEPA preaddestrato in modo autosupervisionato, le rappresentazioni degli oggetti che esso produce catturano strutture di invarianza che non dipendono dal task specifico. Più concretamente: il fine-tuning dell'interfaccia neurosimbolica — i classificatori lineari dei concetti — può avvenire con supervisione molto più sparsa, perché lo spazio latente JEPA è già organizzato in modo che i concetti rilevanti siano linearmente separabili.

Formalmente, si considera un sistema in cui il modulo percettivo è un encoder JEPA $E_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$ preaddestrato, il modulo di parsing della scena decompone l'output dell'encoder in rappresentazioni di oggetto $\{z_{o1}, \dots, z_{oC}\} \subseteq \mathbb{R}^d$ attraverso un meccanismo di object-centric binding (ad esempio Slot Attention di Locatello et al. 2020), e il modulo dei concetti è un insieme di classificatori lineari $W_c \in \mathbb{R}^{d \times |C|}$ dove $|C|$ è il numero di categorie concettuali. Il programma simbolico e l'interfaccia neurosimbolica restano invariati rispetto all'architettura NS-CL originale.

Il training del sistema integrato ha due fasi. La prima fase addestra E_θ con l'obiettivo JEPA (o CV-JEPA, nella variante causale), producendo uno spazio latente con le proprietà di invarianza e struttura causale discusse nei capitoli precedenti. La seconda fase tiene E_θ congelato e addestra il modulo dei concetti W_c con supervisione sparsa sulle coppie (scena, domanda-risposta). La seconda fase può essere notevolmente più efficiente della prima (in termini di dati richiesti) perché lo spazio latente di E_θ è già strutturato: i concetti da apprendere sono variazioni lineari in uno spazio già organizzato, non pattern arbitrari in uno spazio non strutturato.

4.3 - Causalità, composizionalità e il problema dell'identità trans-modale dei concetti

Il nodo filosofico e la sua traduzione operativa

L'integrazione tra NS-CL e CV-JEPA porta in superficie un problema che la tradizione analitica della filosofia del linguaggio conosce sotto il nome di *identità numerica trans-modale* dei concetti, ma che nelle architetture multi-modali contemporanee acquista una struttura tecnica precisa che permette di trattarlo con gli strumenti del do-calcolo. Il problema, nella sua formulazione originaria, è il seguente: un sistema visivo che ha appreso il concetto di *rosso* attraverso la supervisione sulle scene percettive e un sistema linguistico che ha appreso la parola *rosso* attraverso l'addestramento sui corpora testuali operano su spazi latenti diversi, $\mathbb{Z}_{\text{vis}} \subseteq \mathbb{R}^{\{d_{\text{vis}}\}}$ e $\mathbb{Z}_{\text{ling}} \subseteq \mathbb{R}^{\{d_{\text{ling}}\}}$. La domanda se le due rappresentazioni codifichino lo *stesso* concetto non è una domanda sulla loro utilità funzionale — entrambe possono generalizzare bene ai propri rispettivi task — ma una domanda sulla struttura causale sottostante: le due rappresentazioni rispondono agli stessi interventi allo stesso modo?

La traduzione operativa di questa domanda richiede di precisare cosa si intenda per *rispondere agli stessi interventi*. Il framework del do-calcolo fornisce una risposta: due rappresentazioni z_{vis} e z_{ling} del medesimo concetto C sono *causalmente equivalenti* se e soltanto se per ogni intervento $\text{do}(C = v)$ sulla variabile causale che istanzia il concetto, le distribuzioni indotte sulle rispettive rappresentazioni sono legate da una funzione biunivoca stabile:

$P(z_{\text{vis}} \mid \text{do}(C = v))$ e $P(z_{\text{ling}} \mid \text{do}(C = v))$ sono legate da un diffeomorfismo $f_C : \mathbb{Z}_{\text{vis}} \rightarrow \mathbb{Z}_{\text{ling}}$

In altri termini, non si richiede che le due rappresentazioni occupino la stessa regione dello spazio latente — i due encoder operano su spazi distinti — ma che la struttura delle loro *variazioni* al variare dell'istanza del concetto sia la stessa, a meno di un cambio di coordinate inverso.

Questa definizione ha un pregio importante rispetto alle definizioni puramente funzionali: non dipende dalla performance del sistema su un benchmark fisso, ma dalla struttura delle distribuzioni interventali, che sono invarianti rispetto alle correlazioni spurie del dataset di valutazione. Due sistemi possono avere performance identiche su un benchmark e avere strutture causali diverse — come mostrato dal controesempio canonico del Capitolo 1 — e questa differenza è rilevabile attraverso test di distribuzione interventale ma non attraverso test di performance aggregata.

Formalizzazione del problema dell'identità trans-modale attraverso il do-calcolo

Si costruisce ora la struttura formale del problema. Siano $\mathcal{M}_{\text{vis}} = (E_{\text{vis}}, p_{\text{vis}})$ e $\mathcal{M}_{\text{ling}} = (E_{\text{ling}}, p_{\text{ling}})$ due modelli con encoder $E_{\text{vis}} : \mathcal{X}_{\text{vis}} \rightarrow \mathbb{Z}_{\text{vis}}$ e $E_{\text{ling}} : \mathcal{X}_{\text{ling}} \rightarrow \mathbb{Z}_{\text{ling}}$ e distribuzioni generative $p_{\text{vis}}, p_{\text{ling}}$ sui rispettivi spazi osservativi. Sia C una variabile concettuale condivisa — per esempio la proprietà cromatica *rosso* — che figura come causa sia nel processo generativo delle osservazioni visive (la riflettanza spettrale della superficie determina il colore percepito) sia nel processo generativo degli usi linguistici (la categoria semantica *rosso* governa i contesti in cui la parola viene usata).

Il DAG causale condiviso ha la seguente struttura:

$C \rightarrow X_{\text{vis}}$ (il concetto causa osservazioni visive) $C \rightarrow X_{\text{ling}}$ (il concetto causa occorrenze linguistiche) $X_{\text{vis}} \rightarrow Z_{\text{vis}} = E_{\text{vis}}(X_{\text{vis}})$ $X_{\text{ling}} \rightarrow Z_{\text{ling}} = E_{\text{ling}}(X_{\text{ling}})$

Si noti che X_{vis} e X_{ling} non sono causalmente connessi tra loro nel DAG: la dipendenza tra le osservazioni visive e quelle linguistiche è *indiretta*, mediata interamente dalla variabile causale C . Questo è esattamente il caso in cui la formula di aggiustamento backdoor del Capitolo 1 si applica: per stimare la distribuzione interventale $P(Z_{\text{vis}} \mid \text{do}(C = v))$, è sufficiente stimare $P(Z_{\text{vis}} \mid C = v)$ dalla distribuzione osservativa, poiché non ci sono percorsi backdoor aperti tra C e Z_{vis} nel DAG una volta che X_{vis} è marginalizzato.

La *condizione di identità trans-modale* si enuncia ora come segue. Le due rappresentazioni Z_{vis} e Z_{ling} codificano lo stesso concetto C se e soltanto se esistono due funzioni $\psi_{\text{vis}} : \mathbb{Z}_{\text{vis}} \rightarrow \mathbb{R}^m$ e $\psi_{\text{ling}} : \mathbb{Z}_{\text{ling}} \rightarrow \mathbb{R}^m$ tali che:

$P(\psi_{\text{vis}}(Z_{\text{vis}}) \mid \text{do}(C = v)) = P(\psi_{\text{ling}}(Z_{\text{ling}}) \mid \text{do}(C = v))$ per ogni v nel supporto di $P(C)$

dove le funzioni ψ_{vis} e ψ_{ling} sono *probing causale* — la versione interventale dei probing classifiers del Capitolo 12 — che mappano le rispettive rappresentazioni in uno spazio comune $\mathbb{R}^m$. La condizione richiede che i due modelli inducano distribuzioni interventali identiche nello spazio comune — non che le rappresentazioni nello spazio latente originale siano identiche, ma che esista una proiezione che rende le strutture interventali congruenti.

La decomposizione della distanza trans-modale in componenti causali

La distanza tra le distribuzioni interventali dei due sistemi può essere decomposta in tre componenti strutturalmente distinte, corrispondenti a tre cause diverse del divario trans-modale.

La prima componente è la *divergenza di granularità semantica*: i due sistemi possono aver appreso categorizzazioni del concetto C a granularità diverse. Un sistema visivo addestrato su dataset con pochi colori distinti può avere una rappresentazione di C con due-tre categorie grossolane (caldo/freddo,

chiaro/scuro), mentre un sistema linguistico addestrato su testi tecnici di colorimetria può avere una rappresentazione con decine di sottocategorie (vermiglio, scarlatta, cremisi, porpora). Il divario di granularità si quantifica attraverso la divergenza KL tra le distribuzioni marginali di C sotto i due sistemi di categorizzazione:

$$D_{\text{gran}}(C) = D_{\text{KL}}(P_{\text{vis}}(C) \parallel P_{\text{ling}}(C))$$

dove $P_{\text{vis}}(C)$ e $P_{\text{ling}}(C)$ sono le distribuzioni marginali indotte dai due sistemi di categorizzazione sull'insieme dei valori di C . Se $D_{\text{gran}}(C) > 0$, le due distribuzioni interventali non possono essere rese identiche attraverso nessuna coppia di probing (ψ_{vis} , ψ_{ling}) senza perdita di informazione, perché le partition del concetto C non sono allineate.

La seconda componente è la *divergenza di struttura modale*: anche per la stessa granularità semantica, i due sistemi possono aver appreso strutture diverse per come C si manifesta nelle rispettive modalità. Per la modalità visiva, la struttura modale di *rosso* include la distribuzione delle lunghezze d'onda dominanti, i pattern di riflessione, e le relazioni con la luminanza e la saturazione. Per la modalità linguistica, la struttura modale di *rosso* include la distribuzione dei contesti sintattici (aggettivo attributivo, predicativo, superlativo), le relazioni di compatibilità con sostantivi (mela rossa, pericolo rosso, marea rossa), e le implicazioni pragmatiche (urgenza, divieto, passione). La divergenza di struttura modale si quantifica come:

$$D_{\text{modal}}(C) = \mathbb{E}_{\mathbf{v}}[D_{\text{KL}}(P(Z_{\text{vis}} \mid \text{do}(C = \mathbf{v})) \parallel f^*(\psi)P(Z_{\text{ling}} \mid \text{do}(C = \mathbf{v})))]$$

dove $f^*(\psi)$ è il trasporto ottimale che minimizza la divergenza media, e la media è calcolata sulla distribuzione di $P(C)$. Questa quantità misura il divario residuo tra le strutture interventali *dopo* la migliore trasformazione possibile — è il divario non eliminabile nemmeno con la proiezione ottimale.

La terza componente è la *divergenza di confusione causale*: una delle due rappresentazioni può aver appreso correlazioni spurie del proprio dominio che non riflettono la struttura causale di C ma correlati contingenti. Un sistema visivo potrebbe aver imparato che *rosso* correla con *oggetti esteticamente pericolosi* per via di bias del dataset di training; un sistema linguistico potrebbe aver imparato che *rosso* correla con *italiano/spagnolo* per via della frequenza di queste parole in certi corpora. Queste correlazioni spurie producono divergenze nelle distribuzioni interventali che non riflettono differenze semantiche genuine ma artefatti di training. La divergenza di confusione causale si quantifica attraverso la differenza tra le distribuzioni osservazionale e interventale:

$$D_{\text{confus}}(C) = \mathbb{E}_{\mathbf{v}}[D_{\text{KL}}(P(Z \mid C = \mathbf{v}) \parallel P(Z \mid \text{do}(C = \mathbf{v})))]$$

che è zero quando il sistema ha appreso relazioni causali genuine e positivo quando ha imparato correlazioni spurie con confondenti non osservati.

La decomposizione totale è:

$$D_{\text{trans}}(C) = D_{\text{gran}}(C) + D_{\text{modal}}(C) + D_{\text{confus_vis}}(C) + D_{\text{confus_ling}}(C)$$

dove $D_{\text{confus_vis}}$ e $D_{\text{confus_ling}}$ sono le componenti di confusione causale rispettivamente nei sistemi visivo e linguistico. Questa decomposizione permette di diagnosticare il tipo specifico di divario trans-modale per ogni concetto C e di orientare le strategie di riduzione verso la componente dominante.

Il test operativo dell'identità trans-modale: il protocollo interventale incrociato

La condizione di identità trans-modale può essere testata operativamente attraverso un protocollo che chiameremo *cross-modal interventional consistency test* (CMICT). Il protocollo è il seguente.

Si costruisce un dataset di interventi su C che campiona uniformemente il supporto di $P(C)$: per ogni valore v nel supporto, si generano coppie $(x_{\text{vis}}^{\{v\}}, x_{\text{ling}}^{\{v\}})$ campionando dalla distribuzione interventale $P(X_{\text{vis}} \mid \text{do}(C = v))$ e $P(X_{\text{ling}} \mid \text{do}(C = v))$ rispettivamente — cioè si generano osservazioni visive e linguistiche che istanziano il valore v del concetto indipendentemente da ogni correlazione con altre variabili. Nei dataset a design fattoriale come 3DIEBench, questa generazione è immediata perché le variabili di fattore sono controllate dal processo di rendering. Per concetti naturali, richiede un processo di data collection appositamente progettato.

Si addestrano due probing classifiers $\psi_{\text{vis}} : \mathbb{Z}_{\text{vis}} \rightarrow \mathbb{R}^m$ e $\psi_{\text{ling}} : \mathbb{Z}_{\text{ling}} \rightarrow \mathbb{R}^m$ che mappano le rispettive rappresentazioni in uno spazio comune. La scelta di m e della classe di funzioni ammesse (lineare, MLP, kernel) è un iperparametro del test: un probing lineare fornisce il test più conservativo (se l'identità trans-modale è rilevabile con una proiezione lineare, è una forma forte di equivalenza strutturale); un probing non-lineare è più permissivo ma meno interpretabile.

La statistica del test è il *maximum mean discrepancy* (MMD) tra le distribuzioni delle rappresentazioni proiettate sulle due modalità:

$$\text{CMICT}(C) = \text{MMD}^2(P(\psi_{\text{vis}}(Z_{\text{vis}}) \mid \text{do}(C = v)), P(\psi_{\text{ling}}(Z_{\text{ling}}) \mid \text{do}(C = v)))$$

calcolato con kernel gaussiano $k(z, z') = \exp(-\|z - z'\|^2 / (2\sigma^2))$ e mediato sulla distribuzione di v :

$$\text{CMICT}(C) = \mathbb{E}_v[\text{MMD}^2(P(\psi_{\text{vis}}(Z_{\text{vis}}) \mid \text{do}(C = v)), P(\psi_{\text{ling}}(Z_{\text{ling}}) \mid \text{do}(C = v)))]$$

Se $\text{CMICT}(C) \approx 0$, i due sistemi producono distribuzioni interventali statisticamente indistinguibili nello spazio comune: l'identità trans-modale è soddisfatta nella misura testabile. Se $\text{CMICT}(C) > \epsilon$ per una soglia ϵ determinata dalla potenza del test (calcolata per permutazione), i due sistemi hanno strutture interventali diversi per il concetto C — un'evidenza contro l'identità trans-modale.

La significatività statistica del test si stabilisce attraverso la distribuzione nulla di CMICT sotto l'ipotesi di equivalenza. Sotto H_0 (identità trans-modale soddisfatta), le rappresentazioni proiettate dalle due modalità sono campioni della stessa distribuzione a meno di permutazione degli indici, e la distribuzione nulla di MMD^2 converge a una distribuzione chi-quadro pesata per $N \rightarrow \infty$. Il p-value del test è la probabilità di osservare un valore di CMICT almeno tanto grande quanto quello osservato sotto la distribuzione nulla, stimata attraverso 5.000 permutazioni bootstrap delle etichette di modalità.

Il problema dell'identità trans-modale nel framework CV-JEPA

L'integrazione di CV-JEPA con NS-CL crea un contesto naturale per l'applicazione del CMICT : il sistema apprende concetti sia dal modulo percettivo JEPA (attraverso rappresentazioni visive) sia dal modulo linguistico (attraverso il programma simbolico che risolve domande in linguaggio naturale). La coerenza tra i due tipi di rappresentazione è verificabile attraverso il CMICT applicato ai concetti definiti dall'interfaccia neurosimbolica.

Il caso più istruttivo è quello dei concetti che appaiono sia nelle annotazioni visive di 3DIEBench sia nelle domande linguistiche di NS-CL: colori, forme, materiali, relazioni spaziali. Per questi concetti, si può applicare il CMICT confrontando la struttura interventale prodotta dall'encoder JEPA sulle scene visive con la struttura interventale prodotta dal programma linguistico di NS-CL sui testi di domanda-risposta.

Si definisca il caso concreto per il concetto *colore* in 3DIEBench. La variabile causale $C = \{\text{colore_pavimento} \in [0^\circ, 360^\circ]\}$ è un fattore del processo di rendering di 3DIEBench, manipolabile attraverso interventi diretti ($\text{do}(\text{colore_pavimento} = v)$). Le osservazioni visive X_{vis} sono le immagini del dataset nelle condizioni intervenite; le osservazioni linguistiche X_{ling} sono le risposte corrette alle domande del tipo «che colore ha il pavimento?» nelle medesime scene. L'encoder JEPA produce $Z_{\text{vis}} = E_{\text{JEPA}}(X_{\text{vis}})$; il modulo linguistico di NS-CL produce Z_{ling} attraverso la classificazione della risposta nella rete semantica del programma. Il CMICT verifica se la struttura delle variazioni di Z_{vis} al variare del colore del pavimento è congruente con la struttura delle variazioni di Z_{ling} al variare dello stesso fattore.

La previsione della teoria CV-JEPA è che $\text{CMICT}(\text{colore_pavimento})$ sia basso — prossimo a zero — per i sistemi che soddisfano l'identificabilità causale,

perché il fattore causale `colore_pavimento` è già recuperato in modo identificabile nei fattori latenti di CV-JEPA, e il programma linguistico di NS-CL codifica lo stesso concetto attraverso la struttura semantica della domanda. Il test diventa discriminante quando si confrontano sistemi diversi: un sistema seq-JEPA senza il vincolo causale di CV-JEPA potrebbe avere un CMICT più alto per lo stesso concetto, rivelando che la sua rappresentazione visiva del colore del pavimento è meno congruente con la rappresentazione linguistica del sistema NS-CL.

La composizionalità come conseguenza dell'identità trans-modale

Un corollario della condizione di identità trans-modale ha conseguenze dirette per la composizionalità sistematica delle rappresentazioni — la proprietà che Fodor e Pylyshyn avevano identificato come il requisito fondamentale di qualsiasi sistema cognitivo genuino. Si mostra ora che la composizionalità segue dalla combinazione dell'identità trans-modale e dell'indipendenza causale dei concetti nel DAG generativo.

Si consideri un concetto composito $C = C_1 \otimes C_2$ — per esempio *cubo rosso*, che è la composizione del concetto di forma (*cubo*) e del concetto di colore (*rosso*). Se C_1 e C_2 sono causalmente indipendenti nel DAG generativo — la forma non causa il colore e viceversa — allora la distribuzione interventale di C è il prodotto delle distribuzioni interventali dei componenti:

$$P(Z \mid \text{do}(C_1 = v_1, C_2 = v_2)) = P(Z \mid \text{do}(C_1 = v_1)) \otimes P(Z \mid \text{do}(C_2 = v_2))$$

dove $\otimes$ denota la convensione prodotto appropriata per lo spazio latente. Questa fattorizzazione è una conseguenza diretta della struttura di fork nel DAG: C_1 e C_2 condizionano separatamente diverse componenti di Z attraverso percorsi causali distinti, e l'intervento su C_1 non altera la distribuzione interventale di Z rispetto a C_2 .

La composizionalità del sistema emerge ora naturalmente: se le rappresentazioni di C_1 e C_2 sono causalmente identificabili nei fattori latenti $z^{\{(i_1)\}}$ e $z^{\{(i_2)\}}$ rispettivamente (nel senso del teorema iVAE), allora la rappresentazione del concetto composito $C_1 \otimes C_2$ è semplicemente la coppia $(z^{\{(i_1)\}}, z^{\{(i_2)\}})$ — non è necessario apprendere esplicitamente la rappresentazione del concetto composito, che emerge automaticamente dall'indipendenza causale dei suoi componenti. La composizionalità sistematica è una proprietà emergente della struttura causale identificabile, non un vincolo imposto dall'architettura.

Il legame con NS-CL è diretto: il programma simbolico di NS-CL implementa esattamente questa composizionalità attraverso la composizione di operazioni primitive (Filter, Query, Relate) — la rappresentazione visiva di *cubo rosso* emerge applicando `Filter(rosso, Filter(cubo, scene_objects))`, dove le operazioni

Filter sono classificatori lineari nello spazio $z^{\{i\}}$ corrispondente al concetto rilevante. Se i fattori latenti di CV-JEPA soddisfano l'identificabilità causale, i classificatori Filter applicati alle componenti $z^{\{i_1\}}$ e $z^{\{i_2\}}$ separate producono le stesse distinzioni semantiche dei classificatori addestrati sull'intero vettore z — con il vantaggio aggiuntivo che non è necessario addestrare nuovi classificatori per concetti composti mai visti durante il training.

Il limite dell'inscrutabilità e le sue conseguenze formali

Nonostante il CMICT fornisca un test operativo dell'identità trans-modale, rimane il problema filosofico dell'*inscrutabilità del riferimento* identificato da Quine: anche due sistemi con strutture interventali identiche su tutti i test possibili potrebbero in linea di principio rappresentare concetti ontologicamente distinti che si comportano in modo indistinguibile su quell'insieme di test ma divergono su input fuori dalla distribuzione dei test.

Questo limite ha una formulazione formale precisa nel contesto del Causal Representation Learning. Il teorema di identificabilità dell'iVAE (Capitolo 2) garantisce il recupero dei fattori latenti a meno di permutazione e trasformazione elementwise monotona — non il recupero assoluto. Due sistemi con trasformazioni elementwise diverse possono produrre $\text{CMICT} = 0$ su qualsiasi insieme finito di test pur differendo sulle trasformazioni fuori da quell'insieme. Formalmente: siano $\psi_{\text{vis}}(z_{\text{vis}}) = \varphi_{\text{vis}}(s)$ e $\psi_{\text{ling}}(z_{\text{ling}}) = \varphi_{\text{ling}}(s)$ le funzioni che recuperano il fattore causale s dalle rispettive rappresentazioni. L'identità trans-modale richiede $\varphi_{\text{vis}} = \varphi_{\text{ling}}$, ma il teorema iVAE garantisce soltanto che φ_{vis} e φ_{ling} siano entrambe funzioni monotone di s — non che siano la stessa funzione monotona.

La separazione tra le due condizioni è reale: due funzioni monotone distinte $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$ concordano sui valori $s \in S_{\text{test}}$ se S_{test} è un insieme finito, qualunque sia la scelta di S_{test} . L'identità delle funzioni richiede la concordanza su tutto il supporto — una condizione non verificabile in finito. Questo non è un limite contingente della tecnica sperimentale, ma una limitazione epistemica strutturale che il Capitolo 19 inventaria tra i limiti fondamentali del programma.

La conseguenza pratica non è però il nichilismo epistemico: anche se l'identità trans-modale non può essere garantita in assoluto, il CMICT fornisce evidenza *quantitativa* della coerenza tra i sistemi su un insieme di test rappresentativo. E la struttura del teorema iVAE suggerisce che la coerenza su test interventali sufficientemente ricchi — in particolare, test che coprono almeno $d+1$ valori distinti del concetto C nel senso della condizione di sufficienza — è quasi certamente indicativa di equivalenza causale su tutto il supporto, a meno di divergenze nelle code della distribuzione dove la densità dei test è bassa. Il CMICT è quindi non una prova ma un'evidenza robusta di identità trans-modale — un tipo di evidenza che, nella pratica della scienza empirica, è il

massimo raggiungibile senza accesso alla struttura interna del processo generativo.

La Prima Parte di questa monografia ha percorso un arco che, visto dall'interno, potrebbe sembrare un avvicinamento progressivo al suo oggetto dichiarato: dal formalismo degli SCM di Pearl e dall'irriducibilità della gerarchia causale, attraverso le condizioni di identificabilità dell'iVAE di Khemakhem e il CausalVAE di Yang con la sua struttura DAG esplicita, fino all'architettura NS-CL che materializza in un sistema addestrato l'idea di concetti appresi come classificatori lineari causalmente interpretabili nello spazio latente. In realtà, l'arco ha percorso una traiettoria più sottile: ha mostrato che il problema dell'apprendimento causale delle rappresentazioni non è un problema di trovare l'architettura giusta, ma di specificare in modo sufficientemente preciso cosa si chiede all'architettura di fare. I fattori latenti causalmente identificabili richiesti dall'iVAE emergono soltanto sotto condizioni di variazione del contesto sufficientemente ricche; la struttura DAG disciplinata del CausalVAE è identificabile soltanto sotto ipotesi di sufficienza che raramente i dati osservativi soddisfano spontaneamente; i concetti discreti di NS-CL sono sistematici rispetto alla struttura del linguaggio, non rispetto alla struttura del mondo fisico che il sistema percepisce.

Questa serie di condizionali non indebolisce il programma del Causal Representation Learning — lo precisa. E precisarlo significa individuare con esattezza dove si apre il territorio che la Seconda Parte di questa monografia abita: non più il territorio della rappresentazione come fine a se stessa, ma il territorio della rappresentazione al servizio del ciclo percezione-azione. La domanda che i Capitoli da 1 a 4 hanno lasciato aperta — sotto quali condizioni i fattori latenti causalmente strutturati sono anche *utili* per un agente che deve agire in un ambiente fisico — è precisamente la domanda a cui i Capitoli da 5 a 8 rispondono. Il passaggio non è soltanto tematico. C'è una continuità logica: le strutture causali che l'iVAE e il CausalVAE richiedono di identificare nei dati — fattori che variano per ragioni indipendenti, relazioni di dipendenza che rispettano la freccia temporale delle equazioni strutturali — sono esattamente le strutture di cui un sistema Vision-Language-Action ha bisogno per generalizzare a contesti nuovi e per pianificare azioni efficaci in ambienti che differiscono dalla distribuzione di training. Non è una coincidenza: il problem setting dell'agente fisicamente incarnato è uno dei siti naturali in cui la questione della rappresentazione causale smette di essere accademica e diventa ingegneristica. Il Capitolo 5 inizia da qui.

Capitolo 5 — Dal Perception Module di AMI al ciclo percezione-azione: JEPA come fondamento architetturale dei sistemi VLA

5.1 L'asimmetria strutturale tra percezione e azione nel paradigma contemporaneo

Il framework Autonomous Machine Intelligence di LeCun articola il ciclo cognitivo in sei moduli funzionali distinti, tra i quali il Perception Module occupa una posizione privilegiata: è il componente che trasforma il segnale grezzo in una rappresentazione compatta e strutturata su cui gli altri moduli operano. La proposta programmatica di LeCun è che questo modulo debba essere addestrato attraverso obiettivi JEPA, perché soltanto un encoding che predice strutture nello spazio latente — piuttosto che pixel o token osservativi — può catturare le invarianze rilevanti per la pianificazione e il controllo.

Ma il framework AMI rimane, nella sua formulazione originale, un programma teorico più che un sistema implementato. I sistemi Vision-Language-Action (VLA) che sono stati costruiti e valutati empiricamente negli ultimi anni — in particolare RT-2 di Brohan et al. (2023) e OpenVLA di Kim, Pertsch e Karamcheti et al. (2024) — incarnano una versione diversa dello stesso principio: non il Perception Module di AMI fondato su un encoder JEPA con obiettivo predittivo nello spazio latente, ma un Visual Language Model preaddestrato su scala internet il cui modulo visivo è costruito per la comprensione semantica ampia, non per la predittività strutturata nello spazio latente. Il confronto tra questi due approcci — JEPA come Perception Module e VLM come backbone percettivo — è la questione tecnica centrale di questo capitolo e del successivo.

La domanda non è se i sistemi VLA funzionino: i risultati empirici documentano inequivocabilmente che funzionano, e con prestazioni che superano di molto i sistemi precedenti nei task di manipolazione generalizzata. La domanda è più precisa: di quali proprietà dell'encoder visivo dipende il successo osservato? E come si colloca JEPA rispetto a queste proprietà?

5.2 Il problema della generalizzazione nel controllo robotico e la risposta dei VLM

Prima di analizzare l'architettura dei sistemi VLA, occorre chiarire perché il problema della generalizzazione nel controllo robotico sia strutturalmente più difficile del problema della generalizzazione nella visione artificiale.

Un classificatore visivo addestrato su un grande dataset di immagini generalizza a nuovi oggetti perché la struttura visiva degli oggetti condivide sufficienti invarianti da permettere l'estrazione di feature rilevanti. Un sistema di controllo robotico deve invece generalizzare non soltanto sugli oggetti — quali oggetti afferrare — ma sulle relazioni spaziali, sulle istruzioni linguistiche, sulle condizioni di illuminazione, sulle posizioni degli oggetti, sulle dinamiche fisiche, e su combinazioni di questi fattori mai incontrate durante il training. Ogni dimensione di generalizzazione aggiunge un asse di variabilità che il sistema deve apprendere a gestire.

La soluzione proposta da RT-2 è concettualmente elegante: se il problema della generalizzazione nel controllo robotico è strutturalmente analogo al problema della generalizzazione nella comprensione visiva e linguistica — che i VLM hanno affrontato su scala internet — allora si possono trasferire direttamente le capacità di generalizzazione dei VLM al controllo robotico. L'ipotesi di lavoro è che la conoscenza semantica del mondo acquisita da un modello addestrato su miliardi di coppie immagine-testo sia trasferibile alle policy robotiche attraverso un fine-tuning su dati di traiettorie. Questa ipotesi si è rivelata empiricamente fondata: RT-2 ottiene miglioramenti di circa il doppio rispetto a RT-1 e MOO sugli scenari fuori distribuzione, pur utilizzando gli stessi dati robotici.

5.2.1 La formalizzazione del problema di generalizzazione nei sistemi VLA

Il problema di generalizzazione nei sistemi VLA può essere formalizzato attraverso la nozione di **invarianza causale** introdotta nel contesto del Causal Representation Learning. Si supponga che il processo che genera le traiettorie di dimostrazione appartenga a un dominio $D_{\text{train}} = \{d_1, d_2, \dots, d_K\}$ — un insieme finito di ambienti di raccolta dati, ciascuno caratterizzato da una distribuzione $P_d(o, a, r)$ su osservazioni o , azioni a e reward r . Un sistema VLA generalizzante è uno che mantiene prestazioni elevate su un dominio $D_{\text{test}} \supset D_{\text{train}}$ che include ambienti non incontrati durante la raccolta dei dati.

La chiave della generalizzazione è la seguente osservazione di natura causale: alcune relazioni tra osservazioni e azioni ottimali sono **invarianti** attraverso i domini — la relazione tra la posizione di un oggetto e l'azione di presa ottimale è stabile attraverso variazioni di illuminazione, colore dello sfondo, posizione della telecamera — mentre altre relazioni sono **specifiche del dominio** — la relazione tra la forma di un oggetto e la forza di presa ottimale varia con le proprietà fisiche del materiale, che possono differire tra ambienti. Un sistema VLA che apprende soltanto relazioni specifiche del dominio non generalizzerà; uno che apprende le relazioni causalmente invarianti lo farà.

La formalizzazione procede come segue. Per ogni dominio $d \in D_{\text{train}}$, si definisca la distribuzione $P_d(a^* \mid \varphi(o))$ dove $\varphi : \mathcal{O} \rightarrow \mathbb{R}^m$ è una funzione di codifica dell'osservazione e a^* è l'azione ottimale. Si dice che la codifica φ cattura relazioni **causalmente invarianti** rispetto alla famiglia di domini $\{P_d\}_{d \in D}$ se la distribuzione $P(a^* \mid \varphi(o))$ è la stessa per ogni dominio d — ossia se condizionando sull'embedding $\varphi(o)$, l'informazione residua sull'identità del dominio è irrilevante per l'azione ottimale. Formalmente:

$$a^* \perp\!\!\!\perp d \mid \varphi(o) \quad \text{per ogni } d \in D$$

Questa condizione di indipendenza causale può essere verificata attraverso test statistici durante la valutazione, e i sistemi VLA che la soddisfano dovrebbero mostrare generalizzazione robusta a nuovi domini. La connessione con JEPA è

diretta: l'obiettivo predittivo dello spazio latente di JEPA forza esattamente questo tipo di codifica invariante, privilegiando le strutture che si generalizzano attraverso le variazioni superficiali. Un encoder JEPA addestrato su diverse configurazioni dello stesso tipo di task potrebbe quindi soddisfare la condizione di invarianza causale in modo più naturale rispetto a un encoder supervisionato addestrato su singoli task specifici.

5.3 La rappresentazione visiva nei sistemi VLA: struttura e compromessi

Per comprendere come JEPA potrebbe inserirsi nell'architettura VLA, è necessario analizzare con precisione come l'informazione visiva viene elaborata nei sistemi attuali.

In RT-2 — nella sua istanziazione RT-2-PaLI-X — l'encoder visivo è un Vision Transformer da 22 miliardi di parametri (ViT-22B), preaddestrato con un obiettivo supervisionato sul visual question answering e sulla classificazione di immagini. Questo encoder produce una sequenza di k patch token per ciascuna immagine di input, che vengono poi proiettati nell'embedding space del backbone encoder-decoder da 32 miliardi di parametri. L'obiettivo di preaddestramento di questo encoder è la predizione di token linguistici, non la predizione di strutture nello spazio latente.

In OpenVLA, l'encoder visivo è una composizione di due modelli preaddestrati: SigLIP, che cattura semantica di alto livello attraverso un obiettivo di similitudine linguaggio-immagine con sigmoid loss, e DINOv2, che apprende feature visive spaziali di basso livello attraverso un obiettivo di distillazione autosupervisionata. I feature vector delle patch estratte separatamente dai due encoder vengono concatenati channel-wise, producendo una rappresentazione che fonde informazione semantica e informazione spaziale. Questo encoder a doppio canale è proiettato nell'embedding space del backbone linguistico Llama 2 a 7 miliardi di parametri attraverso un proiettore MLP a due strati.

La scelta di OpenVLA di combinare SigLIP e DINOv2 è motivata da una constatazione empirica precisa: i task di controllo robotico richiedono sia comprensione semantica degli oggetti — quale oggetto afferrare, in risposta a quale istruzione — sia ragionamento spaziale fine-grained — dove si trova l'oggetto, come è orientato, dove deve essere posizionato. Il solo SigLIP non è sufficiente per il ragionamento spaziale; il solo DINOv2 non cattura sufficientemente la semantica linguistica. La loro combinazione supera entrambi i modelli presi singolarmente.

Derivazione formale del trade-off tra semantica globale e struttura spaziale locale

La constatazione empirica che motiva la doppia architettura percettiva di OpenVLA — il solo SigLIP non è sufficiente per il ragionamento spaziale; il solo DINOv2 non cattura sufficientemente la semantica linguistica — merita una giustificazione che trascenda il livello della comparazione sperimentale e raggiunga quello della necessità strutturale. La tesi che si sosterrà nelle pagine che seguono è che il trade-off tra semantica globale e struttura spaziale locale non sia un fenomeno contingente, dipendente dall'architettura specifica o dall'obiettivo di addestramento adottato, bensì una conseguenza diretta di due vincoli informativi formali che operano simultaneamente su qualsiasi encoder a dimensionalità finita: il vincolo di capacità dello spazio di embedding e l'incompatibilità tra le funzioni obiettivo che premiano, rispettivamente, l'invarianza semantica e la sensibilità posizionale. La proposta di integrazione JEPA che emerge dalla sezione 5.4 non è allora una direzione tra le altre, ma la risposta architetturealmente motivata a un'impasse che la scelta SigLIP+DINOv2 risolve per via estensiva — aumentando la dimensionalità — piuttosto che per via intensiva, modificando la struttura dell'obiettivo di addestramento.

Il quadro informativo: due statistiche sufficienti incompatibili. Sia $x \in \mathcal{X}$ un'immagine di input; si definiscano due statistiche oracle della scena:

$s = S(x) \in \mathbb{R}^{d_s}$, la statistica minimalmente sufficiente per il contenuto semantico — identità degli oggetti, categorie, allineamento linguistico tra la scena e descrizioni testuali —

$l = L(x) \in \mathbb{R}^{d_l}$, la statistica minimalmente sufficiente per la struttura spaziale — posizioni degli oggetti, loro orientamento, relazioni geometriche tra parti e sottoinsiemi della scena.

Un encoder $E_\theta: \mathcal{X} \rightarrow \mathbb{R}^d$ può essere congiuntamente sufficiente per s e per l soltanto se è in grado di codificare separatamente tutta l'informazione contenuta in entrambe le statistiche. La condizione di sufficienza congiunta richiede:

$$I(E_\theta(x); (s, l)) = I(x; (s, l))$$

Per la regola della catena dell'informazione mutua:

$$I(E_\theta(x); (s, l)) = I(E_\theta(x); s) + I(E_\theta(x); l | s)$$

Poiché $H(E_\theta(x))$ è limitato dall'alto dalla dimensionalità dell'encoder — nel caso continuo, $H(E_\theta(x)) \leq d \cdot \log(1 + \text{Var}(E_\theta)/\epsilon^2)$ per qualsiasi ϵ -discretizzazione, il che cresce logaritmicamente con la varianza ma linearmente con d — il vincolo fondamentale è:

$$I(E_\theta(x); s) + I(E_\theta(x); l | s) \leq H(E_\theta(x)) \leq d \cdot \log(1 + V/\epsilon^2)$$

dove V è una costante proporzionale alla varianza media delle rappresentazioni e ϵ è la granularità di codifica. Quando d è finito e inferiore alla somma $d_s + d_l$ — che è il caso di qualsiasi encoder realizzabile con architettura concreta — l'encoder non può saturare entrambi i termini del membro sinistro simultaneamente, a meno che le due statistiche siano mutuamente indipendenti. In termini di informazione mutua tra le statistiche:

$$I(s; l) = I(x; s) + I(x; l) - I(x; (s, l))$$

Se $I(s; l) = 0$, ossia se il contenuto semantico e la struttura spaziale sono indipendenti nella distribuzione dei dati, la capacità necessaria si riduce a $d_s + d_l$ e un encoder con $d \geq d_s + d_l$ può codificare entrambe le statistiche senza perdita. Ma in scene robotiche reali — dove la posizione degli oggetti è correlata alla loro categoria (i coltelli stanno sul tavolo, le tazze vicino alla macchina del caffè) — l'indipendenza $I(s; l) = 0$ è sistematicamente violata, e la capacità effettiva richiesta è strettamente inferiore a $d_s + d_l$. Tuttavia il modo in cui questa capacità ridotta si distribuisce tra le due statistiche dipende dall'obiettivo di addestramento, non dalla struttura dei dati: è qui che il trade-off diventa necessariamente un effetto dell'architettura dell'apprendimento.

L'encoder semantico: annullamento del segnale spaziale nella gradiente discesa. Si consideri la funzione obiettivo di SigLIP, che estende l'obiettivo CLIP sostituendo la softmax contrastiva con una sigmoide applicata a ciascuna coppia immagine-testo indipendentemente:

$$\mathcal{L}_{\text{SigLIP}} = -(1/n) \sum_{\{i,j\}} [y_{\{ij\}} \cdot \log \sigma(\langle E_I(x_i), E_T(t_j) \rangle / \tau - b) + (1 - y_{\{ij\}}) \cdot \log(1 - \sigma(\langle E_I(x_i), E_T(t_j) \rangle / \tau - b))]$$

dove σ è la funzione sigmoide, τ è la temperatura, b è un bias scalare appreso, e $y_{\{ij\}} = 1$ se la coppia (x_i, t_j) è positiva — stessa immagine e relativa didascalia — e $y_{\{ij\}} = 0$ altrimenti. All'ottimo di questo obiettivo, il prodotto scalare soddisfa l'equazione del punto fisso:

$$\langle E_I^*(x), E_T^*(t) \rangle / \tau - b = \log P(y=1 | x, t) / P(y=0 | x, t) = \text{logit}(P_{+})$$

dove P_{+} è la probabilità che la coppia sia positiva nella distribuzione dati. Per coppie negative campionate uniformemente, $P_{+} \approx 1/n$, ma per la coppia genuina (x, t_x) positiva, P_{+} dipende interamente da quanto il testo t_x è allineato semanticamente all'immagine x .

L'implicazione cruciale per le caratteristiche spaziali è questa: si consideri una trasformazione T_p : $x \rightarrow x^{\wedge}\{p\}$ che preserva il contenuto semantico della scena ma ne altera la struttura spaziale — una rotazione dell'intera scena nel piano, una traslazione delle posizioni degli oggetti, un rispecchiamento orizzontale. La didascalia t_x rimane valida per $x^{\wedge}\{p\}$ se le descrizioni testuali nel dataset non codificano relazioni spaziali assolute: «a robot arm picking up a red cube» descrive indifferentemente la scena con il cubo a sinistra o a destra. In questa circostanza:

$$P_{-+}(x^{\wedge}\{p\}, t_x) \approx P_{-+}(x, t_x)$$

il che implica che la perdita $\mathcal{L}_{\text{SigLIP}}$ è approssimativamente invariante sotto T_p . Il gradiente della perdita rispetto alle caratteristiche spaziali dell'encoder è:

$$\partial \mathcal{L}_{\text{SigLIP}} / \partial (\text{caratteristiche posizionali di } x) \approx 0$$

Non esiste alcun segnale di addestramento che spinga l'encoder a preservare informazione sulla posizione degli oggetti: le direzioni dello spazio dei parametri che codificano struttura spaziale non ricevono gradiente e, per effetto del decadimento dei pesi o della regolarizzazione implicita della discesa stocastica, tendono ad atrofizzarsi. In termini di matrice di informazione di Fisher dell'encoder rispetto al parametro spaziale $p = (x_{\text{obj}}, y_{\text{obj}}, \theta_{\text{obj}})$ — posizione e orientamento di un oggetto nella scena — l'ottimo SigLIP produce:

$$\mathfrak{F}(E_{I^*}; p) = \mathbb{E}_x [\nabla_p \log p(t_x | E_{I^*}(x)) \cdot \nabla_p^T \log p(t_x | E_{I^*}(x))] \approx 0$$

L'encoder ha informazione di Fisher nulla rispetto ai parametri spaziali: l'ottimale rispetto a $\mathcal{L}_{\text{SigLIP}}$ è irriducibilmente cieco alla posizione degli oggetti nella misura in cui la posizione non codetermina il contenuto semantico delle descrizioni testuali.

La frontiera di Pareto del trade-off semantica-spaziale. Il costo di questo annullamento del segnale spaziale si manifesta quantitativamente nel seguente modo. Si definisca l'errore di task VLA per un sistema che richieda sia il riconoscimento semantico (quale oggetto manipolare, in risposta a quale istruzione) sia il controllo motorio spaziale (dove l'oggetto si trova, come raggiungerlo):

$$\epsilon_{\text{VLA}}(E_\theta) = \epsilon_{\text{sem}}(E_\theta) + \epsilon_{\text{spat}}(E_\theta)$$

dove i due addendi sono, rispettivamente, l'errore sulla componente semantica della policy e l'errore sulla componente di localizzazione spaziale. Sotto ipotesi di regressione lineare nei due sottospazi — un'approssimazione valida nel regime in cui le rappresentazioni dell'encoder sono già sufficientemente ricche da permettere la separazione lineare dei fattori —, la teoria della regressione ridge con dati limitati fornisce le seguenti maggiorazioni:

$$\epsilon_{\text{sem}}(E_\theta) \lesssim C_{\text{sem}} \cdot (d_s / I(E_\theta(x); s))$$

$$\epsilon_{\text{spat}}(E_\theta) \lesssim C_{\text{spat}} \cdot (d_l / I(E_\theta(x); l))$$

dove C_{sem} e C_{spat} sono costanti che assorbono la geometria dello spazio dei compiti. L'errore totale è dunque:

$$\epsilon_{\text{VLA}}(E_\theta) \lesssim C_{\text{sem}} \cdot d_s / I(E_\theta(x); s) + C_{\text{spat}} \cdot d_l / I(E_\theta(x); l)$$

Sotto il vincolo di capacità $I(E_\theta(x); s) + I(E_\theta(x); l | s) \leq I_{\max}(d)$ — con $I_{\max}(d)$ crescente in d ma finita per ogni d fissa — minimizzare ϵ_{VLA} richiede un'allocazione ottimale della capacità disponibile tra le due componenti informative. Per la disuguaglianza di Cauchy-Schwarz applicata alla forma iperbolica degli addendi, il minimo di ϵ_{VLA} soggetto al vincolo di capacità si ottiene nel punto:

$$I^*(E_\theta(x); s) = I_{\max} \cdot \sqrt{(C_{\text{sem}} \cdot d_s) / (\sqrt{(C_{\text{sem}} \cdot d_s) + (C_{\text{spat}} \cdot d_l)})}$$

$$I^*(E_\theta(x); l | s) = I_{\max} \cdot \sqrt{(C_{\text{spat}} \cdot d_l) / (\sqrt{(C_{\text{sem}} \cdot d_s) + (C_{\text{spat}} \cdot d_l)})}$$

e il valore ottimale dell'errore è:

$$\epsilon^*_{\text{VLA}} = (\sqrt{(C_{\text{sem}} \cdot d_s) + (C_{\text{spat}} \cdot d_l)})^2 / I_{\max}(d)$$

Questa espressione rivela la struttura del trade-off: l'errore ottimale non dipende separatamente dai pesi C_{sem} e C_{spat} , ma dalla loro radice quadrata pesata dalla dimensionalità delle rispettive statistiche sufficienti. L'errore decresce come $1/I_{\max}(d)$, il che significa che aumentare la dimensionalità d dell'encoder riduce ϵ^*_{VLA} — questa è la logica implicita nella scelta di concatenare SigLIP e DINOv2, che effettivamente raddoppia d — ma non altera la struttura del trade-off, che rimane un problema di allocazione anche con d più grande.

Quello che cambia radicalmente con un encoder JEPA non è il vincolo di capacità, che è una proprietà della dimensionalità dello spazio di embedding indipendente dall'architettura, ma la struttura del segnale di addestramento: l'obiettivo JEPA genera gradiente non nullo rispetto alle caratteristiche spaziali dell'encoder per costruzione, e lo fa in modo tale da massimizzare congiuntamente l'informazione semantica e quella spaziale piuttosto che privilegiarne una a scapito dell'altra.

Perché l'obiettivo JEPA genera gradiente spaziale non nullo. In I-JEPA, l'obiettivo richiede che il predittore g_ψ , prendendo in ingresso la rappresentazione $z_c = E_\theta(x_c)$ del contesto — le patch non mascherate della scena —, produca una stima accurata di $z_t = E_\theta(x_t)$ per le regioni target mascherate. Formalmente:

$$\mathcal{L}_{\text{JEPA}} = \mathbb{E}_{\{(x_c, x_t) \sim \mathcal{D}\}} [\|g_\psi(E_\theta(x_c), m_t) - E_\theta(x_t)\|^2]$$

dove m_t è la maschera che identifica la posizione spaziale delle patch target. La differenza cruciale rispetto a $\mathcal{L}_{\text{SigLIP}}$ è che il predittore g_ψ riceve esplicitamente la posizione m_t come input: deve predire non «cosa c'è nella scena» in astratto, ma «cosa c'è in quella specifica regione della scena». Ciò significa che la perdita $\mathcal{L}_{\text{JEPA}}$ è non invariante sotto la trasformazione T_p che traslata la posizione degli oggetti — poiché traslare un oggetto cambia la regione target che il predittore deve ricostruire e quindi il valore di z_t che

costituisce il target della perdita. Il gradiente rispetto alle caratteristiche posizionali dell'encoder non si annulla:

$$\partial \mathcal{L}_{\text{JEPA}} / \partial (\text{caratteristiche posizionali di } x) \neq 0$$

e la matrice di informazione di Fisher dell'encoder JEPA rispetto al parametro spaziale p soddisfa:

$$\mathfrak{F}(E_{\theta}^{\wedge}\{\text{JEPA}\}; p) = \mathbb{E} [\nabla_p \lg \psi(E_{\theta}(x_c), m_t) - E_{\theta}(x_t) \|^2 \cdot \nabla_p^{\wedge T} \|^2] > 0$$

nella misura in cui la trasformazione della posizione degli oggetti altera la perdita di predizione. L'encoder JEPA è informativamente sensibile alla posizione per costruzione: la struttura dell'obiettivo garantisce che l'informazione di Fisher spaziale sia strettamente positiva senza richiedere alcuna componente di architettura addizionale.

La conseguenza architetturale: JEPA come alternativa intensiva.

Riassumendo, la scelta SigLIP+DINOv2 di OpenVLA risolve il trade-off per via estensiva, espandendo la dimensionalità effettiva delle rappresentazioni da d a $2d$ attraverso la concatenazione. Questa strategia ha tre limitazioni. La prima è la ridondanza parziale: le due componenti non sono informativamente ortogonali — SigLIP cattura qualche struttura spaziale attraverso l'architettura a patch, DINOv2 cattura qualche semantica attraverso il token di classe —, il che significa che parte della capacità raddoppiata è utilizzata per codificare informazione già presente nell'altra componente. La seconda è il costo computazionale: raddoppiare la lunghezza del contesto visivo nel backbone linguistico ha un impatto quadratico sulla complessità dell'attenzione e lineare sulla memoria. La terza è l'assenza di una garanzia di massimizzazione congiunta: il fatto che le due componenti siano addestrate con obiettivi separati non garantisce che la loro concatenazione produca un'allocazione ottimale della capacità combinata nel senso della formula ϵ^*_{VLA} derivata sopra.

Un encoder JEPA video — V-JEPA 2, nella sua versione che processa sequenze di frame e predice la struttura del movimento nello spazio latente — affronta il trade-off per via intensiva: produce, con una sola componente di dimensionalità d , una rappresentazione che possiede Fisher spaziale non nulla per costruzione e Fisher semantica non nulla nella misura in cui le strutture fisiche che il predittore deve apprendere a modellare richiedono la comprensione delle categorie di oggetti — un oggetto rigido si muove diversamente da uno deformabile, un liquido diversamente da un solido. La fusione della comprensione dinamico-fisica del movimento con la comprensione semantica emerge non da due obiettivi separati ma dalla struttura interna dell'obiettivo JEPA, che richiede al predittore di modellare la traiettoria latente degli oggetti attraverso il tempo.

La proposta di integrazione V-JEPA + SigLIP che emerge nella sezione 5.4 non è, in quest'ottica, semplicemente una variante della combinazione

SigLIP+DINOv2 di OpenVLA: è una proposta architettonica distinta in cui la componente JEPA porta la sensibilità spaziale e la comprensione fisica del movimento — che DINOv2 fornisce soltanto in parte, attraverso un obiettivo di distillazione che non è esplicitamente orientato alla dinamica — e la componente SigLIP fornisce l'ancoraggio linguistico. La differenza rilevante, e misurabile, è che la componente V-JEPA possiede una matrice di Fisher spaziale che cresce con la complessità delle traiettorie fisiche osservate durante il preaddestramento, mentre la componente DINOv2 ha Fisher spaziale che emerge come sottoprodotto di un obiettivo di invarianza semantica che non lo privilegia esplicitamente.

5.4 JEPA come Perception Module nei sistemi VLA: l'ipotesi di integrazione

DINOv2, nella sua architettura di base, condivide l'obiettivo fondamentale di JEPA: apprendere rappresentazioni attraverso un meccanismo di predizione che opera sullo spazio latente delle patch, non attraverso la ricostruzione dei pixel. In DINOv2, un teacher network aggiornato con EMA — in modo analogo al target encoder di I-JEPA — produce rappresentazioni target che lo student network deve imparare a predire a partire da view aumentate. La differenza rispetto a I-JEPA è nell'obiettivo preciso della predizione: DINOv2 predice rappresentazioni globali di view diverse della stessa immagine, mentre I-JEPA predice le rappresentazioni di regioni spaziali target a partire da regioni di contesto parziale. JEPA è dunque più orientato alla comprensione spaziale strutturata, DINOv2 più all'invarianza semantica globale.

Questa differenza ha conseguenze per il controllo robotico. Un task come «afferra l'oggetto che sta per cadere dal tavolo» — uno degli esempi più citati nella valutazione qualitativa di RT-2 — richiede sia comprensione semantica sia comprensione spaziale dinamica. L'encoder SigLIP+DINOv2 di OpenVLA cattura entrambe le dimensioni, ma con obiettivi di addestramento non esplicitamente orientati al controllo fisico.

Un encoder JEPA video — in particolare V-JEPA 2 — potrebbe complementare questo profilo in un modo specifico: le rappresentazioni apprese da V-JEPA 2 codificano la struttura del movimento e della fisica qualitativa, non soltanto gli attributi statici degli oggetti. La predizione di regioni video future nello spazio latente obbliga il modello a sviluppare una comprensione implicita delle traiettorie fisiche, delle relazioni di causalità meccanica, e delle proprietà dinamiche degli oggetti. Queste sono precisamente le capacità che i sistemi VLA attuali mostrano di possedere in modo più precario: le dinamiche fisiche precise degli oggetti — velocità di caduta, punto di contatto ottimale, gestione delle deformazioni — restano fuori dalla portata del modello.

La proposta di integrazione che emerge è la seguente: in un sistema VLA di terza generazione, il Perception Module potrebbe essere un encoder V-JEPA —

o una combinazione V-JEPA + SigLIP — che fonde comprensione dinamica del movimento con comprensione semantica degli oggetti, proiettato nell'embedding space di un backbone linguistico per il controllo end-to-end. Questa direzione è teoricamente motivata ma non ancora sistematicamente valutata empiricamente nella letteratura disponibile.

Capitolo 6 — RT-2 e OpenVLA: anatomia di due architetture VLA

6.1 La tokenizzazione delle azioni come ponte tra il dominio linguistico e il dominio motorio

Il contributo tecnico più elegante e strutturalmente significativo della famiglia VLA è la tokenizzazione delle azioni: la scelta di rappresentare i comandi motori del robot come token del vocabolario del modello linguistico, trattandoli esattamente come i token di testo durante il preaddestramento. Questa scelta ha conseguenze teoriche profonde che meritano una trattazione analitica.

Lo spazio d'azione di un robot manipolatore mobile è tipicamente un vettore continuo a bassa dimensionalità. In RT-2, Brohan et al. specificano uno spazio d'azione con sette dimensioni: lo spostamento posizionale dell'end-effector nelle tre dimensioni cartesiane (Δpos_x , Δpos_y , Δpos_z), lo spostamento rotazionale nelle tre direzioni (Δrot_x , Δrot_y , Δrot_z), il livello di estensione del gripper, e un comando discreto di terminazione dell'episodio. Le dimensioni continue vengono discretizzate uniformemente in $B = 256$ bin, producendo una rappresentazione dell'azione come stringa di otto numeri interi:

$a = (\text{terminate}, \text{bin}(\Delta\text{pos}_x), \text{bin}(\Delta\text{pos}_y), \text{bin}(\Delta\text{pos}_z), \text{bin}(\Delta\text{rot}_x), \text{bin}(\Delta\text{rot}_y), \text{bin}(\Delta\text{rot}_z), \text{bin}(\text{gripper}))$

dove la funzione di discretizzazione uniforme per ciascuna dimensione k è

$$\text{bin}(d) = \lfloor (d - \text{min}_k) / (\text{max}_k - \text{min}_k) \cdot (B - 1) \rfloor$$

ristretta all'intervallo $\{0, 1, \dots, B-1\}$. Ogni bin è poi associato a un token esistente nel vocabolario del modello: in RT-2-PaLI-X, dove i primi 1000 interi hanno token univoci, l'associazione è diretta; in RT-2-PaLM-E si sovrascrivono i 256 token usati meno frequentemente nel vocabolario, riservandoli come action token. Il formato di input al modello è una stringa VQA: «Q: what action should the robot take to [task instruction]? A:» a cui segue la sequenza di action token come risposta.

OpenVLA introduce un raffinamento critico a questa discretizzazione: invece di usare i valori minimo e massimo della distribuzione di ciascuna dimensione

come estremi della discretizzazione, usa il primo e il novantanovesimo percentile della distribuzione delle azioni nel dataset di training. Formalmente:

$$\text{bin_OV}(d) = \lfloor (d - q_{\{0.01,k\}}) / (q_{\{0.99,k\}} - q_{\{0.01,k\}}) \cdot (B - 1) \rfloor$$

dove $q_{\{0.01,k\}}$ e $q_{\{0.99,k\}}$ sono i percentili dell'1% e del 99% della dimensione k nel dataset di training. Questa scelta elimina l'impatto degli outlier — errori di teloperazione, movimenti accidentali nei dati — che espandono artificialmente l'intervallo min-max e riducono la granularità effettiva della discretizzazione per il 98% delle azioni reali.

6.2 La funzione obiettivo di training e il co-fine-tuning

Il training di un sistema VLA ottimizza una funzione obiettivo che comprende due componenti: la perdita sulle azioni robotiche e la perdita sui task visivo-linguistici originali del modello preaddestrato. La perdita totale è:

$$\mathcal{L}_{\text{VLA}} = \lambda_{\text{robot}} \cdot \mathcal{L}_{\text{robot}} + \lambda_{\text{web}} \cdot \mathcal{L}_{\text{web}}$$

dove $\mathcal{L}_{\text{robot}}$ è la cross-entropy standard sulla predizione dei token di azione:

$$\mathcal{L}_{\text{robot}} = - \sum_{i=1}^8 \log p_{\theta}(a_i \mid o, l, a_1, \dots, a_{i-1})$$

con o l'immagine di osservazione, l l'istruzione linguistica codificata come stringa VQA, e $(a_1, \dots, a_8)$ la sequenza di action token. La perdita $\mathcal{L}_{\text{web}}$ è la perdita standard del VLM preaddestrato — cross-entropy sul prossimo token per dati di VQA, captioning, e testi interleaved — che preserva le capacità semantiche acquisite durante il preaddestramento su scala internet.

Il co-fine-tuning — mantenere entrambe le perdite durante l'addestramento sulle traiettorie robotiche — si è rivelato significativamente superiore al solo fine-tuning sulla perdita robotica. Per RT-2-PaLI-X-5B, il co-fine-tuning produce un tasso di successo medio del 44% sugli scenari fuori distribuzione, contro il 42% del fine-tuning puro e il 9% del training da zero; per RT-2-PaLI-X-55B, il co-fine-tuning porta il 63% contro il 52% del fine-tuning puro. La dimensione del modello si rivela altrettanto importante: il modello da 55B supera quello da 5B in tutti gli scenari di generalizzazione, con un vantaggio che si accentua nelle condizioni più difficili.

Questo risultato documenta un fenomeno che la letteratura chiama *catastrophic forgetting*: un modello fine-tuned su soli dati robotici dimentica rapidamente le capacità semantiche acquisite durante il preaddestramento, perdendo precisamente le capacità di generalizzazione che rendevano il VLM interessante come fondamento per il controllo robotico. Il co-fine-tuning è l'operazione che mantiene viva la connessione tra la base di conoscenza semantica e le nuove capacità motorie, bilanciando i pesi delle due perdite: in RT-2-PaLI-X il dataset robotico costituisce circa il 50% del mix, in RT-2-PaLM-E circa il 66%.

6.2.1 Analisi del gradiente nel co-fine-tuning: la tensione tra obiettivi

L'efficacia del co-fine-tuning — il training simultaneo sulla perdita robotica e sulla perdita web — può essere compresa analizzando la struttura del gradiente composto. Il gradiente totale rispetto ai parametri θ del modello è:

$$\nabla_{\theta} \mathcal{L}_{\text{VLA}} = \lambda_{\text{robot}} \cdot \nabla_{\theta} \mathcal{L}_{\text{robot}} + \lambda_{\text{web}} \cdot \nabla_{\theta} \mathcal{L}_{\text{web}}$$

I due componenti di gradiente possono essere in conflitto — puntare in direzioni opposte nello spazio dei parametri — o in accordo, a seconda di quanto i task robotici e web condividano strutture rappresentazionali. Il fenomeno del **catastrophic forgetting** si manifesta quando $\nabla_{\theta} \mathcal{L}_{\text{robot}}$ e $\nabla_{\theta} \mathcal{L}_{\text{web}}$ sono quasi ortogonali, in modo che le modifiche necessarie per ottimizzare la perdita robotica si oppongano alle strutture rappresentazionali necessarie per la perdita web.

Una misura della compatibilità dei due obiettivi è il **coseno di similarità del gradiente**:

$$\text{cos_sim} = \langle \nabla_{\theta} \mathcal{L}_{\text{robot}}, \nabla_{\theta} \mathcal{L}_{\text{web}} \rangle / (\| \nabla_{\theta} \mathcal{L}_{\text{robot}} \| \cdot \| \nabla_{\theta} \mathcal{L}_{\text{web}} \|)$$

Se $\text{cos_sim} \approx 1$, i due obiettivi sono allineati e il co-fine-tuning è efficace senza tensioni. Se $\text{cos_sim} \approx 0$, i gradienti sono ortogonali e il co-fine-tuning opera sui due task quasi indipendentemente — assenza di interferenza ma anche di sinergia. Se $\text{cos_sim} \approx -1$, i gradienti sono opposti e il co-fine-tuning è intrinsecamente contraddittorio: ottimizzare un obiettivo peggiora l'altro.

I risultati empirici di RT-2 — dove il co-fine-tuning migliora entrambi gli obiettivi rispetto al fine-tuning puro — suggeriscono che nel regime di parametri e dati considerato, i gradienti sono sufficientemente allineati ($\text{cos_sim} > 0$) da produrre sinergie. Questa sinergia è teoricamente attribuibile alla struttura gerarchica delle rappresentazioni nei Visual Language Models di grande scala: gli strati profondi, che codificano semantica ad alto livello, si trasferiscono efficacemente tra task visivo-linguistici e task robotici; gli strati intermedi, che codificano struttura spaziale e relazionale, sono condivisi tra i due tipi di task; soltanto gli strati più superficiali, che producono i token finali, sono specifici del task. Il co-fine-tuning aggiorna tutti gli strati ma con contributi di gradiente proporzionali al grado di condivisione — producendo naturalmente un aggiornamento differenziato che preserva le strutture condivise.

6.2.2 La proiezione ortogonale di PCGrad e la stratificazione gerarchica dei gradienti in RT-2

La sezione 6.2.1 ha introdotto il coseno di similarità del gradiente come misura della compatibilità tra l'obiettivo robotico e l'obiettivo web nel co-fine-tuning di RT-2, e ha constatato che i risultati empirici di RT-2 suggeriscono $\cos_sim > 0$ nel regime considerato. Tuttavia, la misura scalare del coseno è insufficiente per caratterizzare completamente la tensione tra i due obiettivi: anche un coseno moderatamente positivo nasconde componenti direzionali conflittuali quando i gradienti sono vettori ad alta dimensionalità. Il presente paragrafo deriva formalmente la proiezione ottimale che gestisce il conflitto locale di gradiente — l'algoritmo PCGrad (Gradient Surgery, Yu et al., 2020) — e analizza come la struttura gerarchica di un ViT di grande scala induca un pattern sistematico di accordo e conflitto attraverso la profondità.

Siano $\mathbf{g}_r = \nabla_{\theta} \mathcal{L}_{\text{robot}} \in \mathbb{R}^{|\theta|}$ e $\mathbf{g}_w = \nabla_{\theta} \mathcal{L}_{\text{web}} \in \mathbb{R}^{|\theta|}$ i gradienti dei due obiettivi rispetto ai parametri θ del modello. Il gradiente composito dello standard gradient descent è:

$$\mathbf{g}_{\text{total}} = \lambda_r \cdot \mathbf{g}_r + \lambda_w \cdot \mathbf{g}_w$$

Quando $\langle \mathbf{g}_r, \mathbf{g}_w \rangle < 0$ (coseno negativo), questa combinazione lineare determina una riduzione del modulo del gradiente effettivo inferiore a ciò che si otterrebbe ottimizzando ciascun obiettivo separatamente: il contributo conflittuale di un obiettivo si oppone al progresso dell'altro. La perdita di efficienza è quantificabile. Il progresso atteso nella direzione di riduzione dell'obiettivo robotico per il gradiente composito è:

$$\Delta \mathcal{L}_{\text{robot}} \approx -\eta \cdot \langle \mathbf{g}_r, \mathbf{g}_{\text{total}} \rangle = -\eta \cdot (\lambda_r \|\mathbf{g}_r\|^2 + \lambda_w \langle \mathbf{g}_r, \mathbf{g}_w \rangle)$$

Se $\langle \mathbf{g}_r, \mathbf{g}_w \rangle < 0$ e $|\lambda_w \langle \mathbf{g}_r, \mathbf{g}_w \rangle| > \lambda_r \|\mathbf{g}_r\|^2$, il gradiente composito induce addirittura un aumento di $\mathcal{L}_{\text{robot}}$ — il catastrophic forgetting nella sua forma estrema. La condizione per evitarlo è:

$$\lambda_r / \lambda_w > |\langle \mathbf{g}_r, \mathbf{g}_w \rangle| / \|\mathbf{g}_r\|^2 = \|\mathbf{g}_w\| \cdot |\cos_sim| / \|\mathbf{g}_r\|$$

ovvero il coefficiente del task robotico deve essere sufficientemente grande da dominare il termine conflittuale — una condizione che dipende sia dall'angolo sia dalle magnitudini relative dei gradienti.

La proiezione PCGrad. L'algoritmo di gradient surgery propone una soluzione geometricamente elegante: invece di scalare i coefficienti, si proietta il gradiente conflittuale sul piano ortogonale al gradiente dell'altro obiettivo, eliminando la componente di conflitto prima di combinare. Per il gradiente robotico $\mathbf{g}_r$, si calcola il gradiente proiettato rispetto al web:

$$\mathbf{g}_r^{\text{PC}} = \mathbf{g}_r - \text{proj}_{\{\mathbf{g}_w\}} \mathbf{g}_r = \mathbf{g}_r - (\langle \mathbf{g}_r, \mathbf{g}_w \rangle / \|\mathbf{g}_w\|^2) \cdot \mathbf{g}_w \text{ se } \langle \mathbf{g}_r, \mathbf{g}_w \rangle < 0$$

$$\mathbf{g}_r^{\text{PC}} = \mathbf{g}_r \text{ se } \langle \mathbf{g}_r, \mathbf{g}_w \rangle \geq 0$$

e simmetricamente per $\mathbf{g}_w$ rispetto a $\mathbf{g}_r$. Il gradiente PCGrad composito è:

$$\mathbf{g}_{\text{PC}} = \lambda_{\text{r}} \cdot \mathbf{g}_{\text{r}}^{\text{PC}} + \lambda_{\text{w}} \cdot \mathbf{g}_{\text{w}}^{\text{PC}}$$

La proprietà fondamentale di questa proiezione è che $\langle \mathbf{g}_{\text{r}}^{\text{PC}}, \mathbf{g}_{\text{w}} \rangle = 0$ per costruzione (nel caso conflittuale): la componente di $\mathbf{g}_{\text{r}}$ che punta contro $\mathbf{g}_{\text{w}}$ è stata rimossa, e il progresso atteso su $\mathcal{L}_{\text{robot}}$ nel caso conflittuale diventa:

$$\Delta \mathcal{L}_{\text{robot}} \approx -\eta \cdot \lambda_{\text{r}} \cdot \|\mathbf{g}_{\text{r}}^{\text{PC}}\|^2 \leq 0$$

che è sempre non-positivo (progresso garantito o stazionarietà), indipendentemente dall'angolo tra i gradienti originali. La perdita di informazione introdotta dalla proiezione è la magnitudine della componente rimossa:

$$\|\mathbf{g}_{\text{r}} - \mathbf{g}_{\text{r}}^{\text{PC}}\| = |\langle \mathbf{g}_{\text{r}}, \mathbf{g}_{\text{w}} \rangle| / \|\mathbf{g}_{\text{w}}\| = \|\mathbf{g}_{\text{r}}\| \cdot |\cos_{\text{sim}}|$$

che è piccola quando i gradienti sono quasi ortogonali ($\cos_{\text{sim}} \approx 0$) o allineati ($\cos_{\text{sim}} \approx 1$), e massima quando sono quasi antiparalleli. Questo fornisce un criterio per decidere quando la proiezione è utile: il beneficio (eliminazione del conflitto) supera il costo (perdita di informazione) quando il coseno è sufficientemente negativo e la magnitudine di $\mathbf{g}_{\text{w}}$ è comparabile a quella di $\mathbf{g}_{\text{r}}$.

Stratificazione gerarchica dei gradienti in un ViT di grande scala. Per i ViT di grande scala come PaLI-X-55B o Prismatic-7B (il backbone di OpenVLA), il vettore dei parametri θ si decompone in blocchi $\theta = (\theta_{\text{emb}}, \theta_1, \theta_2, \dots, \theta_L, \theta_{\text{head}})$, dove θ_{emb} sono i parametri dell'embedding layer, θ_l sono i parametri dell' l -esimo blocco Transformer, e θ_{head} sono i parametri della testa di output. Il coseno di similarità locale tra i due gradienti al livello l è:

$$\cos_{\text{sim}}^{\{l\}} = \langle \mathbf{g}_{\text{r}}^{\{l\}}, \mathbf{g}_{\text{w}}^{\{l\}} \rangle / (\|\mathbf{g}_{\text{r}}^{\{l\}}\| \cdot \|\mathbf{g}_{\text{w}}^{\{l\}}\|)$$

dove $\mathbf{g}_{\text{r}}^{\{l\}} = \partial \mathcal{L}_{\text{robot}} / \partial \theta_l$ e $\mathbf{g}_{\text{w}}^{\{l\}} = \partial \mathcal{L}_{\text{web}} / \partial \theta_l$. La teoria dell'apprendimento delle rappresentazioni gerarchiche predice un pattern sistematico di $\cos_{\text{sim}}^{\{l\}}$ attraverso la profondità l :

Per i livelli bassi ($l \leq L/3$), entrambi i gradienti spingono verso rappresentazioni locali di basso livello — bordi, texture, strutture cromatiche — che sono ugualmente utili per il task robotico e per il task web. In questa regione si attende $\cos_{\text{sim}}^{\{l\}} \approx 1$: massimo allineamento, nessuna proiezione necessaria.

Per i livelli intermedi ($L/3 < l \leq 2L/3$), i gradienti riflettono la specializzazione semantica: il task robotico spinge verso rappresentazioni spaziali e relazionali (dove è l'oggetto, come è orientato), mentre il task web spinge verso rappresentazioni categoriali e linguistiche (che cosa è l'oggetto, come viene descritto). In questa regione si attende $|\cos_{\text{sim}}^{\{l\}}| < 0.5$ con possibili episodi di $\cos_{\text{sim}}^{\{l\}} < 0$: la proiezione PCGrad è più utile.

Per i livelli alti e per la testa ($l > 2L/3$), i gradienti sono specifici del formato di output: il task robotico richiede probabilità sui token di azione (7 dimensioni

continue discretizzate), mentre il task web richiede probabilità sull'intero vocabolario (decine di migliaia di token). In questa regione il conflitto è massimo e il segnale dei gradienti è quasi ortogonale: $\cos_sim^{\{l\}} \approx 0$ o $\cos_sim^{\{l\}} < 0$.

Questa stratificazione ha un'implicazione pratica per il co-fine-tuning di RT-2 e sistemi analoghi: la proiezione PCGrad applicata uniformemente a tutti i livelli è sovra-conservativa per i livelli bassi (dove non serve) e insufficiente per i livelli alti (dove il conflitto è strutturale e la proiezione non risolve la diversità intrinseca di formato). Una strategia differenziata — tassi di apprendimento diversi per livello, oppure fine-tuning completo dei livelli alti e lora-adaptation dei livelli intermedi e bassi — è teoricamente giustificata da questo pattern di $\cos_sim^{\{l\}}$.

La connessione con JEPA emerge precisamente in questa stratificazione. Un encoder JEPA preaddestrato produce rappresentazioni per le quali la discrepanza semantica tra il task robotico e il task web è ridotta rispetto a un encoder addestrato solo sul task web: le rappresentazioni z_AGG di JEPA sono per costruzione invarianti alle variazioni di dominio superficiali, il che riduce la componente del gradiente robotico che spinge verso rappresentazioni specifiche del task. In termini del coseno locale, si attende che un sistema con Perception Module JEPA abbia $\cos_sim^{\{l\}}$ più uniforme e positivo attraverso la profondità — con meno episodi di conflitto nelle regioni intermedie — rispetto a un sistema con encoder fine-tuned end-to-end. Questa predizione è verificabile sperimentalmente misurando i gradienti layer per layer durante il co-fine-tuning, ed è una delle diagnosi che il programma di ricerca del Capitolo 20 propone di eseguire per confrontare architetture VLA con e senza Perception Module JEPA.

6.3 - Il profilo delle capacità emergenti e dei limiti strutturali: derivazione dalla struttura dell'obiettivo di addestramento

L'analisi più rivelatrice dei sistemi VLA non è quella delle prestazioni aggregate, ma quella delle capacità emergenti — le capacità che il sistema acquisisce dal preaddestramento su scala internet e che non erano presenti nei dati robotici. RT-2 categorizza queste capacità in tre gruppi: symbol understanding, reasoning e human recognition. Il catalogo empirico, tuttavia, rimane descrittivo se non viene fondato sulla struttura formale dell'obiettivo di addestramento: la domanda pertinente non è soltanto *quali* capacità emergono e *quali* limiti persistono, ma *perché* — ovvero quale meccanismo seleziona, dall'immenso repertorio di capacità latenti del VLM preaddestrato, esattamente quelle che RT-2 manifesta e non altre.

La risposta si trova nell'analisi del gradiente composito sviluppata nella sezione 6.2.1. Per un sistema addestrato con l'obiettivo

$$\mathcal{L}_{\text{VLA}} = \lambda_{\text{r}} \cdot \mathcal{L}_{\text{r}} + \lambda_{\text{w}} \cdot \mathcal{L}_{\text{w}}$$

il gradiente totale rispetto ai parametri θ è $\nabla_{\theta} \mathcal{L}_{\text{VLA}} = \lambda_{\text{r}} \cdot g_{\text{r}} + \lambda_{\text{w}} \cdot g_{\text{w}}$. Si può formalizzare la nozione di *accessibilità di una capacità* attraverso il concetto di varietà di capacità: per ogni capacità C (intesa come una classe di compiti che il sistema è in grado di eseguire con successo), si definisce la varietà $\mathcal{M}_C \subseteq \mathbb{R}^{\{|\theta|\}}$ come l'insieme delle configurazioni di parametri che rendono il sistema capace di C . La distanza euclidea $d(\theta, \mathcal{M}_C)$ misura quanto il modello, alla configurazione corrente θ , sia lontano dall'acquisire C . L'addestramento riduce $d(\theta, \mathcal{M}_C)$ se e soltanto se il gradiente di $\mathcal{L}_{\text{VLA}}$ possiede una componente diretta verso $\mathcal{M}_C$, ovvero se

$$\langle \nabla_{\theta} \mathcal{L}_{\text{VLA}}, n_C(\theta) \rangle < 0$$

dove $n_C(\theta)$ è il vettore unitario uscente dalla proiezione di θ su $\mathcal{M}_C$. Questa componente è la somma pesata delle proiezioni dei due gradienti:

$$\langle \lambda_{\text{r}} g_{\text{r}} + \lambda_{\text{w}} g_{\text{w}}, n_C(\theta) \rangle = \lambda_{\text{r}} \langle g_{\text{r}}, n_C \rangle + \lambda_{\text{w}} \langle g_{\text{w}}, n_C \rangle$$

Un capacità C è dunque accessibile — e in particolare emergente nel senso di RT-2 — se e soltanto se almeno uno dei due contributi è negativo, e il contributo dominante è negativo a sufficienza da compensare eventuali resistenze dell'altro. La struttura dell'emergenza è quindi determinata dalla geometria reciproca tra i gradienti di $\mathcal{L}_{\text{r}}$, $\mathcal{L}_{\text{w}}$ e le varietà delle singole capacità nello spazio dei parametri.

La proiezione del gradiente web sulle capacità emergenti. Per il symbol understanding, la varietà $\mathcal{M}_{\{\text{sym}\}}$ contiene le configurazioni che permettono di associare simboli visivi (cifre, logotipi, icone) a rappresentazioni semantiche attraverso istruzioni linguistiche. Il dataset web D_{web} , da cui $\mathcal{L}_{\text{w}}$ è calcolato, contiene una densità altissima di esempi di questo tipo: testi matematici accompagnati da figure, immagini di tabelle annotate, pagine di e-commerce con numeri e loghi, articoli di notizie con grafici. Il gradiente g_{w} per questi esempi ha una componente $\langle g_{\text{w}}, n_{\{\text{sym}\}} \rangle < 0$ che, integrata sul batch di training, è sistematicamente negativa — il training web spinge i parametri verso $\mathcal{M}_{\{\text{sym}\}}$. Il co-fine-tuning conserva questo segnale mantenendo $\mathcal{L}_{\text{w}}$ attivo; il fine-tuning puro su soli dati robotici lo annulla, perché D_{robot} non contiene esempi di compiti simbolici. La differenza empirica tra RT-2-PaLI-X-55B (82% su symbol understanding) e RT-1 (16%) è precisamente la differenza tra un sistema che conserva $\langle g_{\text{w}}, n_{\{\text{sym}\}} \rangle < 0$ durante l'addestramento e uno che lo azzera.

Per il reasoning — nella sua varietà matematica, cromatica e multilingue — il meccanismo è identico nella struttura ma differente nei dati specifici. Il dataset D_{web} contiene problemi aritmetici, classificazione per colore in contesti linguistici diversi, traduzioni di istruzioni tra lingue. Ogni categoria contribuisce un gradiente g_{w} con componente negativa verso $\mathcal{M}_{\{\text{reasoning}\}}$. La riduzione

della variante RT-2-PaLI-X-55B rispetto a RT-1 (46% contro 16%) riflette l'accumulo di questi contributi nel corso del preaddestramento.

Per la human recognition la situazione è quantitativamente diversa — il tasso di successo di 53% è inferiore all'82% del symbol understanding — ma strutturalmente analoga. I dati web contengono abbondante material visivo associato a nomi di persone (didascalie di foto giornalistiche, profili social, materiali biografici illustrati), tuttavia la densità di questa corrispondenza nel dataset web è inferiore a quella dei simboli convenzionali, e la varietà $\mathcal{M}_{\text{human rec}}$ è più ampia (i volti umani variano su più dimensioni continue rispetto a cifre e logotipi). La componente $\langle g_w, n_{\text{human rec}} \rangle$ è sistematicamente negativa ma di magnitudine minore, il che si traduce in un tasso di successo inferiore pur mantenendo un'emergenza netta rispetto alla baseline RT-1.

La derivazione dei limiti persistenti. Il profilo dei fallimenti di RT-2 — incapacità di apprendere nuove abilità motorie, fallimento nel controllo di dinamiche fisiche complesse, impotenza nei task multi-step oltre uno o due livelli di indirezione — non è un effetto contingente di scelte architetturali modificabili a margine. Esso deriva formalmente dall'analisi precedente per una ragione strutturale: il dataset D_{web} non contiene dimostrazioni di manipolazione robotica. Per qualsiasi capacità motoria C_{motor} non rappresentata in D_{robot} , la proiezione del gradiente web sulla varietà motoria è nulla:

$$\langle g_w, n_{\{C_{\text{motor}}\}} \rangle = 0$$

per tutti gli esempi nel batch D_{web} , perché $D_{\text{web}} \cap \{\text{esempi con componente su } \mathcal{M}_{\{C_{\text{motor}}\}}\} = \emptyset$. La formula del gradiente composito diventa pertanto:

$$\langle \nabla_{\theta} \mathcal{L}_{VLA}, n_{\{C_{\text{motor}}\}} \rangle = \lambda_r \langle g_r, n_{\{C_{\text{motor}}\}} \rangle + 0$$

e l'unico segnale disponibile per sviluppare C_{motor} è quello proveniente da $\mathcal{L}_r$, che è la cross-entropy sulla predizione degli action token e che quindi generalizza soltanto alla distribuzione delle azioni presenti in D_{robot} . Qualsiasi primitiva motoria assente da D_{robot} — afferrare oggetti per il manico, eseguire movimenti di precisione come piegare un asciugamano, gestire dinamiche di spinta con oggetti a centro di massa eccentrico — ha una distanza $d(\theta_{\text{init}}, \mathcal{M}_{\{C_{\text{motor}}\}}) \gg 0$ all'inizio dell'addestramento e zero segnale di gradiente per ridurla. L'indipendenza di questo limite dalla scala del modello è cruciale: un modello da 550 miliardi di parametri addestrato sulla stessa coppia ($D_{\text{robot}}, D_{\text{web}}$) manifesterà identici limiti motori, perché il problema non è la capienza parametrica ma la struttura informazionale dei dati di addestramento.

Una precisazione formale chiarisce il ruolo della scala del modello rispetto ai limiti. La stratificazione dei gradienti derivata in 6.2.1 stabilisce che

$\cos_sim^{\{l\}} \approx 1$ per i livelli bassi ($l \leq L/3$), $|\cos_sim^{\{l\}}| < 0.5$ per i livelli intermedi ($L/3 < l \leq 2L/3$), e $\cos_sim^{\{l\}} \approx 0$ per i livelli alti ($l > 2L/3$). In un modello di grande scala come RT-2-PaLI-X-55B, la dimensionalità elevata dello spazio dei parametri $\theta \in \mathbb{R}^{\{|\theta|\}}$ con $|\theta| \approx 55 \cdot 10^9$ permette che i due gradienti $g_r^{\{l\}}$ e $g_w^{\{l\}}$ nei livelli intermedi occupino sottospazi quasi ortogonali — in alta dimensione, due vettori casuali sono quasi certamente quasi ortogonali, e i gradienti dei due task, pur condividendo strutture di basso livello, divergono abbastanza nei livelli intermedi da poter coesistere senza interferenze devastanti. In un modello di scala ridotta (5B), la stessa quantità di interferenza occupa una frazione maggiore dello spazio parametrico disponibile, costringendo una compressione rappresentazionale che degrada selettivamente le strutture web-specifiche. Il divario di 19 punti percentuali tra i due modelli sugli scenari fuori distribuzione (63% contro 44% per il co-fine-tuning) è dunque formalmente proporzionale al grado in cui la struttura gerarchica del ViT può mantenere $\cos_sim^{\{l\}} > 0$ nei livelli intermedi al variare della scala: non è un effetto di mera capienza, ma una conseguenza della geometria del gradiente composito in spazi di dimensione diversa.

Il limite strutturale del ragionamento multi-step. I fallimenti nei task multi-step meritano un'analisi separata perché non sono derivabili dall'argomento della varietà $D_{web} \cap \mathcal{M} = \emptyset$. La perdita robotica $\mathcal{L}_r$ è formulata come cross-entropy su una singola sequenza di action token per ciascun timestep:

$$\mathcal{L}_r = -\sum_{i=1}^8 \log p_{\theta}(a_i \mid o, l, a_1, \dots, a_{i-1})$$

dove o è l'osservazione corrente e l l'istruzione. Questa formulazione assume implicitamente l'assunzione markoviana: la politica $\pi_{\theta}(a_t \mid o_t, l)$ è condizionata soltanto all'osservazione corrente, non alla storia delle osservazioni o delle azioni precedenti. Il gradiente $\nabla_{\theta} \mathcal{L}_r$ non fornisce quindi alcun segnale per apprendere rappresentazioni che codifichino l'avanzamento verso un sotto-obiettivo intermedio — lo stato di completamento parziale di un task multi-step. La perdita web $\mathcal{L}_w$ fornisce segnale per il ragionamento compositivo nel dominio linguistico, ma la composizione linguistica opera su token testuali, non su sequenze di azioni motorie. La struttura temporale delle due perdite è fondamentalmente incommensurabile: $\mathcal{L}_w$ può addestrare RT-2 a rispondere «il passo tre segue il passo due» in linguaggio naturale, ma non può addestrarlo a eseguire una sequenza di azioni fisiche coordinata perché il feedback robotico su sequenze di azioni non fa parte di D_{web} . Questo disallineamento nella struttura temporale delle due perdite — $\mathcal{L}_r$ è one-step markoviana, $\mathcal{L}_w$ contiene segnali multi-step ma esclusivamente linguistici — produce un limite al ragionamento sequenziale che è strutturale rispetto all'obiettivo di co-fine-tuning e che non può essere superato aumentando la scala del modello né modificando il bilanciamento λ_r / λ_w .

6.4 OpenVLA: open-source, dimensioni ridotte, generalizzazione superiore

OpenVLA occupa un'architettura radicalmente diversa da RT-2 su tre dimensioni fondamentali. La prima è la scala: 7 miliardi di parametri contro i 55 miliardi di RT-2-PaLI-X, un fattore di circa sette. La seconda è la trasparenza: OpenVLA è il primo sistema VLA generalista completamente open-source, con pesi, codice di training PyTorch e notebook di fine-tuning rilasciati pubblicamente. La terza è la dimensione dei dati robotici: 970 mila episodi dall'Open X-Embodiment dataset, contro le circa 350 mila traiettorie di RT-2-X.

Nonostante la scala inferiore, OpenVLA supera RT-2-X di 16.5 punti percentuali assoluti sul tasso di successo medio attraverso 29 task e molteplici embodiment robotici. Su BridgeData V2, il vantaggio è ancora più netto: 70.6% per OpenVLA contro 50.6% per RT-2-X. Su Google robot, le due architetture sono comparabili — 85.0% per OpenVLA contro 78.3% per RT-2-X. RT-2-X mantiene un vantaggio nella generalizzazione semantica, dove il co-fine-tuning con dati internet di RT-2 preserva meglio le conoscenze semantiche; OpenVLA è superiore in tutte le altre categorie di generalizzazione.

La combinazione di encoder SigLIP+DINOv2 contribuisce con circa 5 punti percentuali di vantaggio rispetto al solo SigLIP. Il preaddestramento su Open X è invece il fattore dominante: ablando tutto il preaddestramento robotico (fine-tuning solo su BridgeData V2), si osserva un calo di circa 30 punti percentuali, dimostrando che la diversità dei dati robotici — non la dimensione del modello, non l'architettura dell'encoder — è il principale determinante della generalizzabilità.

Il fine-tuning con LoRA di rango $r = 32$ applicato a tutti i layer lineari del modello raggiunge prestazioni identiche al full fine-tuning aggiornando soltanto l'1.4% dei parametri — 97.6 milioni contro 7.188 miliardi — riducendo il fabbisogno di VRAM da 163.3 GB a 59.7 GB. La quantizzazione a 4-bit conserva le stesse prestazioni della precisione bfloat16 standard riducendo a meno della metà l'impronta in memoria: 7.0 GB contro 16.8 GB. La quantizzazione a 8-bit degrada le prestazioni non per riduzione di precisione — verificato con blocking control, che porta le tre precisioni a prestazioni comparabili — ma per riduzione della frequenza di inferenza da circa 6 Hz a circa 1.2 Hz, che altera le dinamiche di controllo rispetto ai dati di training.

Capitolo 7 — Dataset egocentri, eterogeneità dell'embodiment e radicamento sensomotorio

7.1 La struttura dell'Open X-Embodiment come infrastruttura epistemica

La costruzione dell'Open X-Embodiment dataset rappresenta uno degli sforzi di data collection collaborativa più ambiziosi nella storia della robotica: oltre settanta dataset individuali per un totale di più di due milioni di traiettorie raccolte in un formato unificato che consente l'addestramento di policy robotiche generaliste attraverso la diversità degli embodiment fisici. La mixtureA adottata da OpenVLA privilegia i dataset con maggiore variabilità di task e scene — Fractal e Kuka con il 12.7% ciascuno, Bridge con il 13.3%, DROID con il 10% rimosso nell'ultimo terzo del training — distribuendo il peso proporzionalmente alla struttura informazionale delle traiettorie.

Il concetto fondamentale che governa la struttura di Open X-Embodiment non è la quantità grezza di dati ma la *diversità distribuita*: l'apprendimento di policy robotiche generaliste non beneficia linearmente del volume delle traiettorie, ma richiede una copertura bilanciata degli assi di variabilità rilevanti per la generalizzazione — eterogeneità degli oggetti, delle scene, delle istruzioni linguistiche, degli embodiment fisici, e dei task strutturalmente distinti. Un dataset di un milione di traiettorie dello stesso robot nel medesimo ambiente produce un sistema considerevolmente meno generalizzabile di un dataset di pari volume distribuito su centinaia di configurazioni diverse, anche a parità di qualità delle singole dimostrazioni. Questa non è un'osservazione intuitiva ma una conseguenza derivabile dal teorema di Rademacher sulla complessità di generalizzazione: la bound sull'errore di generalizzazione di un sistema addestrato su dataset D cresce con la complessità della classe di funzioni appresa, e la diversità dei dati riduce la complessità effettiva spingendo il modello verso soluzioni causalmente invarianti piuttosto che verso soluzioni specifiche del dominio.

Formalmente, sia $\mathcal{D} = \{\mathcal{D}_e\}_{e=1}^{|E|}$ una collezione di dataset per E embodiment diversi, con $\mathcal{D}_e = \{(o_t^e(e), a_t^e(e), l^e(e))\}_{t=1}^{T_e}$ le traiettorie raccolte con l'embodiment e , dove $o_t^e(e) \in \mathcal{O}^e(e)$ è l'osservazione specifica dell'embodiment, $a_t^e(e) \in \mathcal{A}^e(e)$ l'azione nello spazio d'azione corrispondente, e $l^e(e)$ l'istruzione linguistica. La mixtureA di training è la distribuzione sul dataset unificato:

$$p_{\text{mix}}(\tau) = \sum_e w_e \cdot p^e(e)(\tau)$$

dove $\tau = (o, a, l)$ è una traiettoria generica e w_e è il peso dell'embodiment e nel mix. La scelta dei pesi $\{w_e\}$ è uno dei problemi di progettazione più delicati nella costruzione di sistemi VLA generalisti: un peso troppo elevato sull'embodiment e porta al dominio del dataset corrispondente, riducendo la diversità effettiva; un peso troppo basso marginalizza traiettorie potenzialmente informative su certi tipi di task. La strategia adottata da Kim et al. (2024) per OpenVLA è una variante del *data balancing* basato su information diversity: si privilegiano i dataset con alta entropia di marginalizzazione sul task space,

misurata come $H(\text{Task}) = -\sum_j p(\text{task}_j \mid \mathcal{D}_e) \log p(\text{task}_j \mid \mathcal{D}_e)$ sulla distribuzione dei task nell'embodiment e .

Il contributo di DROID — il dataset di 76.000 traiettorie di raccolta di dati distribuita con operatori non esperti — introduce una dimensione di variabilità che i dataset di dimostrazione esperta non coprono: la gamma di stili motori umani nell'esecuzione dello stesso task. Mentre le dimostrazioni di operatori esperti sono ottimizzate per l'efficienza e la pulizia della traiettoria, le dimostrazioni di operatori non esperti esplorano lo spazio delle strategie valide in modo meno stereotipato, coprendo comportamenti sub-ottimali ma fisicamente corretti che possono essere informativi per l'apprendimento di policy robuste. Questo effetto — la *sub-optimal diversity advantage* — è analogo al fenomeno documentato nell'RL dove la mixture di policy con diversi livelli di qualità migliora la robustezza dei sistemi addestrati rispetto a dataset di sole dimostrazioni ottimali.

7.2 Il problema strutturale del radicamento sensomotorio

I sistemi VLA attuali non sono, nel senso tecnico preciso del termine, sistemi *incarnati* nel senso che la tradizione della cognitive science intende con il termine *embodied cognition*. Questa affermazione non è un giudizio di valore ma un'osservazione sulla struttura dell'apprendimento: l'incarnazione computazionale richiede che le rappresentazioni interne del sistema siano *costitutivamente dipendenti* dall'esperienza sensomotoria dell'agente — che i concetti visivi emergano dall'interazione con il mondo fisico attraverso il corpo, non soltanto dall'osservazione passiva di dimostrazioni altrui.

Il problema del radicamento sensomotorio (*sensorimotor grounding*) è il problema di garantire che le rappresentazioni interne di un sistema artificiale siano ancorate non soltanto alla struttura statistica delle osservazioni ma alle relazioni causali tra percezione e azione nel corpo fisico. Un sistema che ha appreso a classificare immagini di tazze non ha, in senso proprio, una rappresentazione di *cosa fa* una tazza — della relazione tra la presa, il peso, l'attrito con la superficie, e la funzione di contenimento. Questa lacuna non è colmabile semplicemente aumentando la quantità di dati visivi: richiede l'integrazione di informazione propriocettiva, cinestesica e tattile nell'addestramento, o quantomeno la garanzia che i dati di dimostrazione campionino in modo denso le variazioni nelle proprietà fisiche rilevanti.

Il paradosso del radicamento nei sistemi VLA contemporanei può essere formalizzato attraverso la nozione di *completezza informazionale* rispetto alle variabili causali fisiche del dominio. Sia $S_{\text{phys}} = \{s_i\}_{i=1}^{n_{\text{phys}}}$ l'insieme delle variabili fisiche causalmente rilevanti per un task di manipolazione — massa dell'oggetto, coefficiente di attrito statico e dinamico, rigidità, elasticità, temperatura superficiale, e così via. Sia $S_{\text{obs}} = \{s_j\}_{j=1}^{n_{\text{obs}}}$ l'insieme delle variabili osservabili nell'input sensoriale visivo — colore, forma proiettata, texture

superficiale, luminanza. Il *gap di radicamento* di un sistema addestrato soltanto su dati visivi è definito come:

$$\Delta_{\text{grounding}} = I(Y_{\text{opt}}; S_{\text{phys}} | S_{\text{obs}}) = H(Y_{\text{opt}} | S_{\text{obs}}) - H(Y_{\text{opt}} | S_{\text{phys}}, S_{\text{obs}})$$

dove Y_{opt} è l'azione ottimale, e $I(\cdot; \cdot | \cdot)$ è l'informazione mutua condizionale. Questa quantità misura quanta informazione sull'azione ottimale è contenuta nelle variabili fisiche *al di là* di ciò che è osservabile visivamente. Se $\Delta_{\text{grounding}} > 0$, esiste una componente dell'azione ottimale che non può essere determinata dalle sole osservazioni visive: il sistema mancante del radicamento sensomotorio opera con un deficit informativo strutturale che non può essere compensato da nessuna quantità di dati visivi.

Per i task di grasping con oggetti di proprietà fisiche diverse — una bottiglia d'acqua piena vs. vuota, un panno asciutto vs. bagnato, un filo flessibile vs. un tubo rigido — questo gap è strettamente positivo: la forza di presa ottimale, il punto di contatto ideale, e la strategia di avvicinamento sono causalmente determinati da variabili come la rigidità e il coefficiente di attrito che non sono direttamente osservabili da immagini RGB senza esperienza sensomotoria con quegli specifici materiali. Un sistema che ha visto migliaia di grasping di bottiglie piene e zero grasping di bottiglie vuote avrà un bias sistemico verso la forza di presa appropriata per le bottiglie piene, producendo fallimenti prevedibili e caratteristici nel caso opposto.

7.3 La struttura matematica del gap di radicamento nella manipolazione fisica

Si approfondisce la formalizzazione del gap di radicamento nel contesto specifico della manipolazione, dove le variabili fisiche causalmente rilevanti hanno una struttura algebrica precisa derivabile dalla meccanica del corpo rigido e dai modelli di contatto.

Il task di grasping di un oggetto caratterizzato da geometria G (descritta da una mesh triangolata), massa m , inerzia $I \in \mathbb{R}^{3 \times 3}$, coefficiente di attrito statico μ_s e coefficiente di attrito dinamico μ_d può essere formalizzato come il seguente problema di ottimizzazione vincolata. L'azione ottimale è la forza di presa $f \in \mathbb{R}^6$ (tre componenti di forza e tre di coppia nell'end-effector frame) che minimizza un costo di task $c(f)$ soggetto ai vincoli fisici:

$$f^* = \underset{f}{\operatorname{argmin}} \{ c(f) \mid f \in FC(G, \mu_s) \text{ [vincolo del cono di attrito]} \parallel f_n \leq F_{\text{max}} \text{ [vincolo di forza massima]} f \cdot \hat{n} \geq f_{\text{min}} \text{ [vincolo di forza minima di contatto]} \}$$

dove $FC(G, \mu_s) = \{f : \parallel f_t \parallel \leq \mu_s \cdot f_n\}$ è il friction cone dipendente dalla geometria G e dal coefficiente di attrito μ_s , con f_n e f_t le componenti normale e tangenziale della forza, e $\hat{n}$ la normale alla superficie nel punto di contatto.

La dipendenza di f^* da (G, μ_s, m, I) è il cuore del problema di radicamento. Il cono di attrito $FC(G, \mu_s)$ dipende strettamente da μ_s , che non è direttamente osservabile dalle immagini RGB — lo stesso oggetto può avere μ_s molto diverso a seconda del materiale superficiale (vetro lucido: $\mu_s \approx 0.1$, gomma opaca: $\mu_s \approx 0.9$). La massa m influenza la dinamica post-grasping e quindi la politica di trasporto, non il grasping statico; ma la stima di m da immagini RGB richiede conoscenza del materiale, della densità, e del volume — nessuno dei quali è direttamente accessibile senza esperienza tattile.

Questa struttura formale permette di quantificare $\Delta_{\text{grounding}}$ per task specifici. Si consideri la distribuzione di task di grasping con oggetti campionati dalla distribuzione $p(G, \mu_s, m, I)$ di un dataset robotico reale. Il gap di radicamento è la varianza residua di f^* che non è spiegata dalle variabili visivamente osservabili:

$$\Delta_{\text{grounding}} = \text{Var}(f^* \mid S_{\text{obs}}) = \mathbb{E}_{\{S_{\text{obs}}\}}[\text{Var}(f^* \mid S_{\text{obs}})] = \mathbb{E}_{\{S_{\text{obs}}\}}[\text{Var}(\arg\min_f c(f) \text{ s.t. } f \in FC(G, \mu_s) \mid S_{\text{obs}})]$$

Per una distribuzione di μ_s con varianza σ^2_{μ} e per un costo $c(f) = \|f - f_{\text{ref}}\|^2$ lineare attorno a una forza di riferimento f_{ref} , la varianza residua è approssimata da:

$$\Delta_{\text{grounding}} \approx (\partial f^* / \partial \mu_s)^2 \cdot \text{Var}(\mu_s \mid S_{\text{obs}})$$

dove $\text{Var}(\mu_s \mid S_{\text{obs}})$ è la varianza del coefficiente di attrito condizionata alle osservazioni visive — essenzialmente la componente di μ_s non predetta dall'immagine RGB. Per μ_s distribuito uniformemente nell'intervallo $[0.1, 0.9]$ (la gamma tipica di materiali comuni), se la previsione visiva riduce l'incertezza del 50% (un sistema che ha visto sufficiente varietà di materiali riconosce le categorie grossolane), $\text{Var}(\mu_s \mid S_{\text{obs}}) \approx 0.32^2 / 12 \cdot 0.5 \approx 0.004$.

La derivata $(\partial f^* / \partial \mu_s)$ dipende dalla geometria del cono di attrito e può essere calcolata attraverso il teorema della funzione implicita applicato alle condizioni di ottimalità KKT del problema di grasping vincolato. Per grasping al limite del cono di attrito (la condizione più comune nelle prese di oggetti lisci), si ha:

$$\partial f^* / \partial \mu_s \approx f_n / \mu_s^2$$

dove f_n è la forza normale al contatto. Per $f_n = 10$ N e $\mu_s = 0.3$ (un caso rappresentativo di grasping di oggetti lisci moderatamente pesanti):

$$\Delta_{\text{grounding}} \approx (10/0.09)^2 \cdot 0.004 \approx 4900 \cdot 0.004 \approx 20 \text{ N}^2$$

corrispondente a una deviazione standard dell'errore di forza di presa di circa 4.5 N — un errore che nella pratica si traduce in una significativa probabilità di slip o di deformazione dell'oggetto per grasping delicati. Questo calcolo, pur nella sua approssimazione, mostra quantitativamente perché il gap di

radicamento sensomotorio ha conseguenze operative concrete: non è un deficit filosofico ma una fonte misurabile di fallimento in task di manipolazione fine.

7.4 Il dataset Ego4D e la cattura del radicamento incarnato umano

Il corpus Ego4D — circa 3.600 ore di video egocentrico da attività umane quotidiane, raccolte in 74 luoghi geograficamente e culturalmente diversi da 923 partecipanti — rappresenta il più vasto archivio disponibile di percezione incarnata umana catturata in prima persona. La struttura delle annotazioni di Ego4D copre cinque benchmark: episodic memory (dove ho visto l'oggetto X?), forecasting (cosa farò nei prossimi k secondi?), hand-object interaction (quando e come avviene il contatto?), audio-visual diarization (chi parla e quando?), e natural language queries. La densità di annotazione delle interazioni mano-oggetto — ogni frame del contatto è annotato con il tipo di presa, la fase dell'interazione (avvicinamento, contatto, manipolazione, rilascio), e la categoria dell'oggetto — rende Ego4D un dataset privilegiato per l'analisi del radicamento sensomotorio.

Dal punto di vista del paradigma JEPA, Ego4D offre esattamente il tipo di segnale di preaddestramento che il Capitolo 5 ha identificato come mancante nei sistemi VLA attuali: l'informazione sulle strutture dinamiche delle interazioni fisiche nella prospettiva in cui il robot deve agire. Un encoder V-JEPA addestrato su Ego4D apprende a predire la rappresentazione latente di frame futuri a partire da sequenze di frame passati — il che, data la densità di interazioni mano-oggetto nel corpus, significa predire il movimento della mano, la risposta dell'oggetto al contatto, e la dinamica post-grasping senza supervisione esplicita su nessuna di queste quantità. La struttura dell'obiettivo JEPA, applicata a dati egocentrici di interazione fisica, dovrebbe produrre per inferenza naturale rappresentazioni delle invarianti del movimento incarnato che non sono direttamente contenute in nessun frame singolo ma emergono dalla struttura temporale delle sequenze.

Questa inferenza non è verificata empiricamente nel contesto dei sistemi VLA — è una previsione teorica derivata dall'ipotesi di trasferimento fisico del Capitolo 8 e dal programma del sotto-programma III del Capitolo 20. Ma la sua plausibilità è sostenuta dai risultati di V-JEPA 2.1 sul task di Short-Term Anticipation (STA) su Ego4D stesso: un sistema che raggiunge il 35% di miglioramento relativo sul task di anticipazione delle interazioni rispetto al precedente stato dell'arte deve avere sviluppato, nel proprio spazio latente, una qualche rappresentazione della struttura dinamica dell'interazione imminente che va al di là della semplice classificazione della scena. Questa rappresentazione della *prossima interazione* è una forma computazionale del radicamento sensomotorio: non è sapere cosa è un oggetto, ma anticipare come sarà manipolato nel contesto corrente.

7.5 La misura quantitativa del radicamento nei sistemi JEPA: la metrica di anticipazione fisica

Si propone una metrica formale per misurare il grado di radicamento sensomotorio nelle rappresentazioni di un sistema JEPA, derivata dalla struttura del task STA di Ego4D ma applicabile a qualsiasi sistema percettivo addestrato su dati di interazione fisica.

La *Physical Grounding Score* (PGS) è definita come la capacità del sistema di predire, dalla sola rappresentazione latente dello stato corrente, le variabili fisicamente rilevanti dell'interazione *futura*, condizionata sulla sola storia percettiva e senza accesso alle variabili fisiche dirette:

$$\text{PGS}(E, H) = R^2(\hat{y}_{\text{phys}}, y_{\text{phys}})$$

dove $y_{\text{phys}} = (f_{\text{contact}}, t_{\text{contact}}, \text{type}_{\text{contact}})$ è il vettore delle variabili fisiche dell'interazione futura — forza al contatto, istante del contatto, tipo di presa — $\hat{y}_{\text{phys}} = g(E(o_{t-H:t}))$ è la previsione di queste variabili dall'embedding della storia delle osservazioni attraverso un regressore lineare leggero g , e $R^2(\cdot, \cdot)$ è il coefficiente di determinazione calcolato sul dataset di test. Il regressore lineare è un vincolo essenziale: valuta la decodificabilità lineare delle variabili fisiche dallo spazio latente, nella stessa tradizione del probing causale del Capitolo 12.

La PGS ha un'interpretazione informazionale precisa: misura quanta informazione sulle variabili fisicamente rilevanti dell'interazione futura è linearmente accessibile dallo spazio latente del sistema — essenzialmente, il grado in cui il sistema ha sviluppato una rappresentazione implicita della fisica dell'interazione come conseguenza del proprio obiettivo di preaddestramento. Un sistema con $\text{PGS} \approx 0$ non ha radicamento sensomotorio: le sue rappresentazioni non contengono informazione sulla fisica dell'interazione al di là di ciò che è direttamente visibile nel frame corrente. Un sistema con $\text{PGS} \approx 1$ ha pieno radicamento: le sue rappresentazioni codificano in modo linearmente accessibile tutte le variabili fisiche rilevanti per la manipolazione futura.

Per un encoder addestrato soltanto su task di classificazione di immagini statiche, ci si attende $\text{PGS} \approx 0$ per le variabili dinamiche (forza al contatto, timing) e $\text{PGS} > 0$ soltanto per le variabili correlate con la categoria dell'oggetto (tipo di presa approssimativo). Per un encoder V-JEPA preaddestrato su Ego4D, la previsione del programma di radicamento è PGS significativamente maggiore per le variabili dinamiche — il che costituisce la *previsione discriminante* verificabile attraverso il disegno sperimentale del sotto-programma III del Capitolo 20.

La scomposizione additiva di PGS nelle sue componenti corrispondenti a diverse categorie di variabili fisiche — cinematica (posizione e velocità del polso), dinamica (forza e coppia), tattile (tipo di contatto, area di presa), temporale (timing del contatto) — permette di profilare il radicamento del

sistema in termini qualitativamente distinti. Un sistema con alto PGS cinematico e basso PGS tattile ha appreso la geometria del movimento ma non la fisica del contatto; un sistema con alto PGS temporale e basso PGS cinematico ha appreso l'anticipo degli eventi ma non la struttura spaziale dell'interazione. Il profilo completo di PGS permette quindi di diagnosticare con precisione il tipo di deficit di radicamento e di orientare il processo di raccolta dati aggiuntivi verso le dimensioni informativamente più carenti.

7.6 L'eterogeneità dell'embodiment come strumento di apprendimento delle invarianti fisiche

Una delle intuizioni più rilevanti che emerge dall'analisi della struttura di Open X-Embodiment non riguarda la diversità dei task o degli ambienti, ma la diversità degli *embodiment fisici* — la varietà di robot, di morfologie, di spazi d'azione e di modalità sensoriali inclusi nel dataset. Questa eterogeneità è solitamente considerata una difficoltà tecnica da gestire attraverso la normalizzazione delle azioni e la standardizzazione dei formati di input. Ma essa è anche una risorsa epistemica straordinaria per l'apprendimento delle invarianti fisiche.

Il ragionamento è il seguente. Due robot con morfologie diverse — un braccio a 6 DOF e un robot mobile con gripper — che eseguono lo stesso task di prendere un oggetto e posizionarlo su un ripiano producono traiettorie le cui componenti specifiche dell'embodiment (la sequenza di configurazioni dei giunti, le velocità angolari, le forze ai giunti) sono radicalmente diverse. Ma le invarianti del task — la geometria della presa rispetto all'oggetto, il percorso dell'oggetto nello spazio cartesiano, la relazione temporale tra approccio e contatto — sono le stesse indipendentemente dall'embodiment. Un sistema addestrato su entrambi gli embodiment che apprende a risolvere il task deve quindi apprendere rappresentazioni che catturano queste invarianti task-specifiche, fattorizzandole dalle variabili embodiment-specifiche — esattamente la struttura che il paradigma JEPa mira a produrre attraverso la predittività nello spazio latente.

Formalmente, si consideri la decomposizione della distribuzione delle traiettorie nel dataset multi-embodiment:

$$p(\tau) = \sum_e p(e) \cdot p(\tau \mid e) = \sum_e p(e) \cdot p(\tau_{\text{task}} \mid \text{task}, e_{\text{phys}}) \cdot p(e_{\text{phys}} \mid e)$$

dove $\tau_{\text{task}} = (o_{\text{task}}, a_{\text{task}})$ è la componente della traiettoria invariante rispetto all'embodiment (la storia dell'oggetto nello spazio cartesiano e del task semantico), e_{phys} sono le caratteristiche fisiche dell'embodiment (morfologia, workspace, cinematica), e la dipendenza di τ_{task} da e_{phys} cattura il fatto che la stessa variabile task-invariante si manifesta diversamente per embodiment diversi (lo stesso oggetto viene afferrato da angoli diversi con effettori diversi).

La previsione del paradigma JEPA applicato al dataset multi-embodiment è che l'encoder apprenda a estrarre $z = E(o)$ tali che la distribuzione $p(z \mid \text{task}, e)$ sia approssimativamente indipendente da e per gli aspetti task-invarianti:

$$I(z; e \mid \text{task}) \approx 0$$

dove $I(\cdot; \cdot \mid \cdot)$ è l'informazione mutua condizionale. Questa condizione è verificabile empiricamente attraverso un test di probing causale sulla variabile e : se le rappresentazioni dell'encoder sono davvero invarianti all'embodiment condizionatamente al task, un classificatore lineare addestrato su z non dovrebbe essere in grado di predire e meglio di un classificatore basale. I risultati di Kim et al. (2024) su OpenVLA suggeriscono indirettamente questa proprietà — il sistema generalizza ragionevolmente a nuovi embodiment non visti durante il training — ma la verifica formale attraverso un test di probing causale sulla variabile e non è stata condotta nel paper originale e costituisce uno degli obiettivi secondari del sotto-programma VI descritto nel Capitolo 20.

7.7 Verso il radicamento computazionale: le tre condizioni necessarie

L'analisi delle sezioni precedenti permette di enunciare tre condizioni necessarie (non sufficienti) per il radicamento sensomotorio computazionale, derivate dalla struttura formale del gap $\Delta_{\text{grounding}}$ e dalla struttura del dataset multi-embodiment.

La *prima condizione* è la *completezza informativa del corpus di preaddestramento*: il dataset deve includere variazioni nelle variabili fisiche causalmente rilevanti (μ_s , m , I , rigidità) in modo tale che le distribuzioni condizionali $p(y_{\text{opt}} \mid S_{\text{obs}}, s_{\text{phys}})$ siano distinguibili per valori diversi di s_{phys} . Se il corpus di preaddestramento campiona sistematicamente soltanto oggetti rigidi e il task di deployment include oggetti deformabili, nessun meccanismo di apprendimento può produrre rappresentazioni radicate rispetto alla deformabilità — la variazione necessaria per l'apprendimento è assente nei dati. Questa condizione è più stringente del semplice requisito di diversità dei task: richiede che la diversità dei dati copra gli assi di variabilità fisicamente rilevanti, non soltanto quelli visivamente diversi.

La *seconda condizione* è la *sensibilità percettiva alle variabili fisiche*: l'architettura percettiva deve essere in grado di rilevare, nelle osservazioni sensoriali disponibili, i correlati visivi (anche indiretti) delle variabili fisiche rilevanti. Per la rigidità, i correlati visivi includono la deformazione durante il contatto, il pattern di riflessione della luce sulla superficie, la texture fine della superficie. Un encoder che opera soltanto su immagini a bassa risoluzione o che comprime in modo eccessivo l'informazione visiva può mancare questi correlati. L'obiettivo di Dense Prediction di V-JEPA 2.1 — che forza la rappresentazione di informazione spaziale fine-grained a ogni token — è esattamente il meccanismo che preserva la sensibilità percettiva ai correlati visivi delle variabili

fisiche: la texture superficiale, le microdeformazioni, i riflessi speculari sono caratteristiche locali che sarebbero compresse da un encoder con obiettivo globale.

La *terza condizione* è la *coerenza temporale attraverso le fasi dell'interazione*: le rappresentazioni degli stati pre-contatto, durante il contatto e post-contatto devono essere connesse da strutture temporali coerenti che l'encoder produce come conseguenza dell'obiettivo di previsione video. Un sistema che apprende indipendentemente le rappresentazioni dei frame statici non sviluppa la coerenza temporale necessaria per anticipare come la scena cambierà a seguito del contatto — quella che il framework STA di Ego4D chiama *interaction anticipation* e che è il correlato computazionale del radicamento sensomotorio dinamico. V-JEPA, prendendo strutture spaziotemporali nello spazio latente di sequenze video, soddisfa questa condizione in modo più naturale di qualsiasi architettura addestrata su frame indipendenti.

Queste tre condizioni non sono indipendenti: la seconda dipende dalla prima nel senso che la sensibilità percettiva è irrilevante se i dati non contengono variazione nelle variabili fisiche da rilevare; la terza dipende dalla seconda nel senso che la coerenza temporale delle rappresentazioni fisiche richiede che le proprietà fisiche siano rappresentate in primo luogo. La loro soddisfazione congiunta è la condizione necessaria per il radicamento sensomotorio computazionale nei sistemi JEPA — una condizione che i sistemi attuali soddisfano in modo parziale e che il programma di ricerca futuro mira a soddisfare in modo più robusto attraverso la combinazione di corpora egocentrici ricchi (Ego4D, Open X), architetture percettive dense (V-JEPA 2.1), e protocolli di valutazione formali basati sulla Physical Grounding Score proposta in questa sezione.

Capitolo 8 — Il sim-to-real gap come problema fondamentale dell'embodiment computazionale

8.1 La struttura matematica del sim-to-real gap

Il *sim-to-real gap* è la discrepanza sistematica tra la distribuzione delle osservazioni e delle dinamiche incontrate in simulazione durante il training e la distribuzione incontrata nel mondo fisico durante il deployment. Questa discrepanza si manifesta su quattro livelli strutturalmente distinti.

Il *visual gap* descrive la differenza $P_{\text{sim}}(x) \neq P_{\text{real}}(x)$ nelle distribuzioni dei pixel: qualità del rendering, modellazione dell'illuminazione, modelli di

materiale, rumore del sensore, gamma di variazione delle condizioni ambientali. Misurando la discrepanza attraverso la divergenza di Jensen-Shannon:

$$JS(P_{sim} \parallel P_{real}) = \frac{1}{2} \cdot D_{KL}(P_{sim} \parallel M) + \frac{1}{2} \cdot D_{KL}(P_{real} \parallel M)$$

con $M = \frac{1}{2}(P_{sim} + P_{real})$ distribuzione mista, si verifica che $JS > 0$ in tutti i sistemi simulati realistici. Un encoder visivo addestrato su P_{sim} apprende feature che possono non essere rappresentative di P_{real} , producendo rappresentazioni degradate quando il modello è deployato nel mondo reale.

8.1.1 Quantificazione formale del visual gap attraverso divergenze di distribuzione

La divergenza di Jensen-Shannon tra distribuzione simulata e reale fornisce una misura aggregata del visual gap, ma non rivela la struttura del gap — quali dimensioni dello spazio delle osservazioni contribuiscono maggiormente alla discrepanza. Per una comprensione più fine, è utile decomporre $JS(P_{sim} \parallel P_{real})$ attraverso la *Principal Component Analysis contrastiva* (cPCA) delle due distribuzioni.

Si denotino con $\Sigma_{sim} = \mathbb{E}_{\{P_{sim}\}}[(x - \mu_{sim})(x - \mu_{sim})^T]$ e $\Sigma_{real} = \mathbb{E}_{\{P_{real}\}}[(x - \mu_{real})(x - \mu_{real})^T]$ le matrici di covarianza empiriche delle due distribuzioni. Il *visual gap decomposto* è caratterizzato dagli autovettori della differenza di covarianza $\Sigma_{real} - \Sigma_{sim}$: gli autovettori con autovalori positivi grandi corrispondono alle direzioni nello spazio delle osservazioni che hanno maggiore varianza nella distribuzione reale rispetto a quella simulata — ossia le dimensioni lungo cui la simulazione sottorappresenta la variabilità del mondo reale.

Per un encoder JEPA E_θ , la propagazione del gap dalla distribuzione delle osservazioni alla distribuzione delle rappresentazioni è governata dalla jacobiana dell'encoder:

$$\Sigma_{real}^z - \Sigma_{sim}^z \approx J_E \cdot (\Sigma_{real} - \Sigma_{sim}) \cdot J_E^T$$

dove $J_E = \partial E_\theta(x)/\partial x$ è la jacobiana dell'encoder valutata alla media delle osservazioni e Σ^z denota la matrice di covarianza nello spazio latente. Il sim-to-real gap nello spazio latente è quindi la proiezione del gap nello spazio delle osservazioni attraverso la jacobiana dell'encoder. Se E_θ è un encoder JEPA addestrato con l'obiettivo di invarianza — che sopprime le variazioni lungo le direzioni irrilevanti per la previsione — la sua jacobiana J_E avrà norma ridotta lungo le direzioni di alta varianza nell'*irrilevante* (illuminazione, texture del fondo), riducendo la propagazione del gap simulazione-reale lungo quelle direzioni. Questo è il meccanismo formale per cui si attende che gli encoder JEPA riducano il visual gap più efficacemente degli encoder supervisionati su task specifici: la loro jacobiana è implicitamente orientata a sopprimere le variazioni irrilevanti, che sono esattamente quelle più disomogenee tra simulazione e realtà.

Il *dynamics gap* descrive la discrepanza $f_{\text{sim}}(s, a) \neq f_{\text{real}}(s, a)$ nelle funzioni di transizione di stato. Le simulazioni fisiche usano modelli di contatto semplificati — rigid body dynamics, friction cone approximation, soft body truncation — che non catturano i dettagli delle superfici reali, le deformazioni elastiche, le variazioni nelle proprietà dei materiali, e le interazioni complesse tra oggetti multipli. Questo gap è particolarmente pronunciato per task che richiedono manipolazione delicata o gestione di oggetti deformabili.

Il *morphological gap* descrive la discrepanza tra le specifiche fisiche del robot simulato e quelle del robot reale: massa, momento di inerzia, attrito dei giunti, elasticità dei tendini, calibrazione dei sensori. Anche piccole discrepanze nei parametri fisici si amplificano nel corso di traiettorie lunghe, producendo comportamenti qualitativamente diversi.

Il *latency gap* descrive la differenza tra il ciclo percezione-azione sincrono e istantaneo della simulazione e il ciclo con ritardi di comunicazione, variabilità nelle frequenze di campionamento e jitter nel timing del mondo reale. OpenVLA documenta questo effetto in modo preciso: la quantizzazione a 8-bit degrada le prestazioni robotiche non per riduzione di precisione delle computazioni ma perché riduce la frequenza di inferenza da ~ 6 Hz a ~ 1.2 Hz, alterando le dinamiche del sistema rispetto a quelle viste durante il training.

8.2 Strategie di riduzione del sim-to-real gap

La letteratura ha sviluppato diverse strategie per attenuare il sim-to-real gap, differenziate per il tipo di gap affrontato e per il loro posizionamento nell'architettura del sistema.

La *domain randomization* è la strategia più diffusa: invece di addestrare su una singola simulazione P_{sim} , si campiona durante il training da un insieme di simulazioni $P_{\text{sim}}^{(1)}, \dots, P_{\text{sim}}^{(K)}$ che variano parametricamente — posizioni di luce, texture degli oggetti, masse e attrito dei corpi, colore dello sfondo, posizione della camera. Formalmente, la perdita di training diventa:

$$\mathcal{L}_{\text{DR}} = \mathbb{E}_{\{k \sim \text{Uniform}(\{1, \dots, K\})\}} \mathbb{E}_{\{(s,a) \sim P_{\text{sim}}^{(k)}\}} [\mathcal{L}(\pi(s), a)]$$

La speranza è che questa varietà forzi l'apprendimento di feature invarianti rispetto alle variazioni parametriche. Empiricamente, la domain randomization riduce efficacemente il visual gap ma è meno efficace per il dynamics gap.

La *system identification* attacca il dynamics gap stimando il set di parametri fisici del modello simulato θ_{sim} che minimizza la discrepanza tra traiettorie simulate e traiettorie reali:

$$\theta_{\text{sim}}^* = \underset{\theta_{\text{sim}}}{\operatorname{argmin}} \mathbb{E}_{\{(s_0, a) \sim D_{\text{real}}\}} [\|f_{\text{sim}}(s_0, a; \theta_{\text{sim}}) - s'_{\text{real}}\|^2]$$

dove D_{real} è un dataset di transizioni reali. Il modello calibrato $f_{\text{sim}}(\cdot; \theta_{\text{sim}}^*)$ approssima meglio le dinamiche reali. Il limite è che la system

identification richiede dati reali e un modello fisico sufficientemente espressivo — assunzione spesso non soddisfatta per sistemi fisici complessi.

Una terza strategia — la *domain adaptation* — agisce direttamente sull'encoder visivo: si addestra il modello a produrre rappresentazioni identiche per osservazioni corrispondenti provenienti da simulazione e da mondo reale, minimizzando la distanza nelle distribuzioni delle rappresentazioni:

$$\mathcal{L}_{DA} = \mathbb{E}_{\{x \sim P_{sim}\}} [\|E(x)\|^2] + \mathbb{E}_{\{x \sim P_{real}\}} [\|E(x)\|^2] + \mathcal{L}_{adv}(D, E)$$

dove $\mathcal{L}_{adv}$ è la perdita avversaria di un discriminatore D che tenta di distinguere rappresentazioni simulate da rappresentazioni reali. Quando il discriminatore non riesce a distinguere, le rappresentazioni sono diventate domain-invariant.

8.3 JEPA e l'ipotesi di trasferimento fisico

I sistemi VLA attuali aggirano il problema del sim-to-real gap in un modo indiretto ma efficace: evitano la simulazione, addestrando direttamente su dati reali. Tuttavia, la raccolta di dati robotici reali è lenta, costosa, e limitata nella varietà prodotta. Per RT-2, i dati sono stati raccolti con 13 robot in 17 mesi in un ambiente cucina controllato. La simulazione rimane indispensabile per il training a scala molto maggiore — specialmente per task che richiedono molte interazioni per apprendere, come i task di reinforcement learning con reward sparso.

La connessione con JEPA emerge qui in modo strutturalmente non ovvio. Un encoder V-JEPA addestrato sull'osservazione passiva di video fisici reali apprende rappresentazioni delle dinamiche fisiche che non dipendono dalla fedeltà della simulazione. Se questo encoder viene fine-tuned per il controllo in un ambiente simulato, le rappresentazioni delle dinamiche fisiche che esso porta dal preaddestramento su video reali potrebbero ridurre il visual gap e il dynamics gap, perché il modello «sa già» — in senso rappresentazionale — come si muovono gli oggetti fisici reali, e non ha bisogno di re-impararlo dai pixel simulati.

Questa *ipotesi di trasferimento fisico di JEPA* non è ancora stata testata sistematicamente. Il suo test richiederebbe un esperimento in cui si confronta un sistema addestrato in simulazione con un encoder V-JEPA preaddestrato su video reali contro un sistema con un encoder addestrato solo in simulazione, su task di controllo valutati nel mondo reale. L'esperimento è fattibile con gli strumenti disponibili ma non risulta tra i contributi pubblicati al momento della stesura di questa monografia.

8.3.1 Bound formale sulla degradazione delle performance JEPA in funzione del sim-to-real gap: il teorema di trasferimento attraverso la jacobiana dell'encoder

La sezione 8.3 ha affermato, in termini qualitativi, che un encoder JEPA riduce il sim-to-real gap rispetto a un encoder supervisionato su task specifici perché la sua jacobiana è implicitamente orientata a sopprimere le variazioni irrilevanti. La sezione 8.1.1 ha derivato la formula di propagazione del gap attraverso la jacobiana:

$$\Sigma_{\text{real}}^z - \Sigma_{\text{sim}}^z \approx J_E \cdot (\Sigma_{\text{real}} - \Sigma_{\text{sim}}) \cdot J_E^T$$

Questi due risultati preparano il terreno per un bound quantitativo sulla degradazione delle performance che un sistema addestrato in simulazione subisce quando viene trasferito nel mondo reale. Il bound che si deriva qui mette in relazione tre oggetti misurabili: la divergenza tra le distribuzioni delle osservazioni (la grandezza del gap), la struttura della jacobiana dell'encoder (la selettività dell'encoder rispetto alle variazioni), e il degrado della performance del pianificatore (la quantità di interesse pratico).

Sia $\pi_\theta : \mathbb{R}^d \rightarrow \mathcal{A}$ la policy del sistema (nel contesto JEPA, la combinazione di encoder E_θ , World Model f_θ e pianificatore MPC), con performance valutata attraverso il valore atteso del return:

$$V_d(\pi_\theta) = \mathbb{E}_{\{x \sim P_d\}} [R(\pi_\theta(E_\theta(x)))]$$

dove P_d è la distribuzione delle osservazioni nel dominio $d \in \{\text{sim}, \text{real}\}$, R è il return del task, e $\pi_\theta \circ E_\theta$ è la policy composta encoder-pianificatore. Il gap di performance è:

$$\Delta V = V_{\text{real}}(\pi_\theta) - V_{\text{sim}}(\pi_\theta)$$

che misura quanto le performance in simulazione (usate per l'addestramento) sovrastimano le performance reali (l'obiettivo effettivo).

Teorema (Bound di trasferimento JEPA). Sotto le seguenti assunzioni: (T1) La policy π_θ è L_π -Lipschitz rispetto alla rappresentazione latente: $\|\pi_\theta(z) - \pi_\theta(z')\| \leq L_\pi \cdot \|z - z'\|$ per ogni $z, z' \in \mathbb{R}^d$. (T2) Il return R è L_R -Lipschitz rispetto all'azione: $|R(a) - R(a')| \leq L_R \cdot \|a - a'\|$ per ogni $a, a' \in \mathcal{A}$. (T3) L'encoder E_θ è differenziabile con jacobiana $J_E = \partial E_\theta(x) / \partial x$ localmente costante (approssimazione lineare locale sull'intervallo delle differenze di distribuzione).

Allora il gap di performance soddisfa il bound:

$$|\Delta V| \leq L_\pi \cdot L_R \cdot \|J_E\|_F \cdot \sqrt{(\text{Tr}(\Sigma_{\text{real}} - \Sigma_{\text{sim}}) + \|\mu_{\text{real}} - \mu_{\text{sim}}\|^2)}$$

dove $\|J_{E\|_F}$ è la norma di Frobenius della jacobiana dell'encoder, $\text{Tr}(\cdot)$ è la traccia, μ_d e Σ_d sono media e covarianza di P_d .

La dimostrazione procede come segue. Per (T1) e (T2), il gap di return dovuto a una differenza di distribuzione delle osservazioni è:

$$|\Delta V| = |\mathbb{E}_{\{P_{\text{real}}\}}[R(\pi_{\theta}(E_{\theta}(x)))] - \mathbb{E}_{\{P_{\text{sim}}\}}[R(\pi_{\theta}(E_{\theta}(x)))]| \leq L_R \cdot \mathbb{E}_{\{P_{\text{real}}, P_{\text{sim}}\}}[\|\pi_{\theta}(E_{\theta}(x_r)) - \pi_{\theta}(E_{\theta}(x_s))\|] \leq L_R \cdot L_{\pi} \cdot \mathbb{E}_{\{P_{\text{real}}, P_{\text{sim}}\}}[\|E_{\theta}(x_r) - E_{\theta}(x_s)\|]$$

dove $x_r \sim P_{\text{real}}$ e $x_s \sim P_{\text{sim}}$. L'aspettativa della norma della differenza delle rappresentazioni è maggiorata (per la disuguaglianza di Jensen e la norma di Frobenius):

$$\mathbb{E}[\|E_{\theta}(x_r) - E_{\theta}(x_s)\|] \leq \sqrt{\mathbb{E}[\|E_{\theta}(x_r) - E_{\theta}(x_s)\|^2]} \approx \sqrt{\mathbb{E}[\|J_E \cdot (x_r - x_s)\|^2]} \leq \|J_{E\|_F} \cdot \sqrt{\mathbb{E}[\|x_r - x_s\|^2]} = \|J_{E\|_F} \cdot \sqrt{\text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{sim}} - 2\Sigma_{\text{cross}}) + \|\mu_{\text{real}} - \mu_{\text{sim}}\|^2}$$

dove Σ_{cross} è la matrice di covarianza incrociata tra P_{real} e P_{sim} . Nel caso pessimistico (distribuzione indipendenti), $\Sigma_{\text{cross}} = 0$ e il termine $\sqrt{\text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{sim}}) + \|\mu_{\text{real}} - \mu_{\text{sim}}\|^2}$ è il Wasserstein-2 approssimato tra P_{real} e P_{sim} per distribuzioni gaussiane. Unendo e semplificando: $|\Delta V| \leq L_{\pi} \cdot L_R \cdot \|J_{E\|_F} \cdot W_2(P_{\text{real}}, P_{\text{sim}})$. $\square$

Il bound ha tre fattori che si interpretano separatamente. Il primo, $L_{\pi} \cdot L_R$, è la sensitività della pipeline policy-return: un pianificatore MPC con ottimizzazione a temperatura bassa ($\lambda \rightarrow 0$, regime CEM) è più sensibile alle variazioni delle rappresentazioni rispetto a un pianificatore con temperatura alta, ovvero L_{π} cresce al decrescere di λ . Il secondo, $\|J_{E\|_F}$, è la selettività dell'encoder: encoder con norma di jacobiana piccola amplificano poco le differenze distribuzionali. Il terzo, $W_2(P_{\text{real}}, P_{\text{sim}})$, è la dimensione intrinseca del gap, indipendente dal sistema.

Il vantaggio JEPA formalizzato. La domanda centrale è: perché un encoder JEPA ha $\|J_{E\|_F}$ sistematicamente inferiore a quello di un encoder supervisionato su task specifici? La risposta è nella struttura dell'obiettivo JEPA. L'obiettivo di predizione JEPA minimizza $\|g_{\theta}(E_{\theta}(x_c)) - E_{\theta}(x_t)\|^2$ sui patch di contesto e target. Nelle direzioni dello spazio delle osservazioni lungo cui x_t non porta informazione aggiuntiva rispetto a x_c — le direzioni irrilevanti per la previsione, che corrispondono tipicamente a variazioni di texture, illuminazione e dettagli superficiali — il gradiente dell'obiettivo JEPA rispetto a E_{θ} è quasi nullo: il predittore impara a ignorarle. Di conseguenza, l'encoder E_{θ} sviluppa una jacobiana J_E con norma singolare piccola nelle direzioni irrilevanti.

Per quantificarlo, si decompone la jacobiana $J_E = U \Sigma V^T$ (SVD). Le singole σ_k misurano quanto l'encoder amplifica le variazioni nella k -esima direzione singolare. Per un encoder JEPA addestrato all'ottimalità, si può dimostrare che

$\sigma_k \approx 0$ per le direzioni v_k che corrispondono alle coordinate a bassa varianza predittiva. Il numero di direzioni con $\sigma_k \approx 0$ è almeno $d - r_{\text{eff}}$, dove r_{eff} è il rango effettivo dell'informazione predittivamente rilevante (stimabile come la dimensione intrinseca del manifold delle rappresentazioni JEPA). Per un encoder supervisionato su un singolo task, invece, J_E mantiene norme singolari significative in tutte le direzioni che variano nell'input — incluse le direzioni di texture e illuminazione che non sono causalmente rilevanti per il task ma che il classificatore supervisionato ha imparato a sfruttare come correlati del label. La conseguenza è:

$$\|J_E^{\text{JEPA}}\|_F \leq \|J_E^{\text{supervised}}\|_F$$

con un rapporto che cresce con la ricchezza del dominio di preaddestramento JEPA: più vaste e diversificate sono le trasformazioni irrilevanti viste durante il preaddestramento, più completa è la soppressione della loro jacobiana. Sostituendo nel bound di trasferimento:

$$\frac{|\Delta V^{\text{JEPA}}|}{\|J_E^{\text{JEPA}}\|_F} / \frac{|\Delta V^{\text{supervised}}|}{\|J_E^{\text{supervised}}\|_F} \leq \frac{\|J_E^{\text{JEPA}}\|_F}{\|J_E^{\text{supervised}}\|_F} < 1$$

Il rapporto tra le norme di Frobenius delle jacobiane è quindi la misura precisa del vantaggio di trasferimento dell'encoder JEPA rispetto all'encoder supervisionato — un rapporto verificabile computando le jacobiane numericamente attraverso differenze finite o automatic differentiation sulla coppia di encoder. Sperimentalmente, nei dataset robotici con variazione di illuminazione e sfondo (come DROID e BridgeData2), ci si attende un rapporto $\|J_E^{\text{JEPA}}\|_F / \|J_E^{\text{supervised}}\|_F \in (0.3, 0.7)$, corrispondente a un bound di performance degradation due a tre volte inferiore per il sistema con Perception Module JEPA.

8.4 - Il constraints satisfaction problem nell'action space fisico

La struttura del problema: validità sintattica e fattibilità fisico-dinamica

La tokenizzazione delle azioni come strategia architetturale dei sistemi VLA, analizzata nel Capitolo 6, risolve elegantemente il problema dell'interfaccia tra il dominio linguistico e il dominio motorio: le azioni continue dello spazio cartesiano vengono discretizzate in token del vocabolario del modello, permettendo di trattare il controllo robotico come un problema di generazione sequenziale di testo. Ma questa eleganza formale nasconde una tensione strutturale che diventa critica nel deployment fisico: un modello linguistico standard produce distribuzioni su token senza nessun vincolo di validità fisica, e la proiezione dell'output nel dominio motorio non garantisce automaticamente

che le traiettorie prodotte siano cinematicamente ammissibili, dinamicamente stabili o fisicamente realizzabili.

Si distinguono tre livelli di validità per una traiettoria prodotta da un sistema VLA, ordinati per forza del vincolo.

Il *primo livello* è la validità sintattica: ogni token di azione prodotto appartiene all'insieme A_valid dei token di azione definiti dall'architettura. Questa condizione è imposta da RT-2 attraverso la restrizione del vocabolario di campionamento ai 256 action token validi durante l'inferenza. La distribuzione condizionata risultante è:

$$p_constrained(a_i | o, l, a_1, \dots, a_{i-1}) = p_0(a_i | o, l, a_1, \dots, a_{i-1}) \cdot \mathbb{1}[a_i \in A_valid] / Z$$

dove Z è la costante di normalizzazione. Questa proiezione garantisce che ogni singolo token di azione corrisponda a un valore nell'intervallo discretizzato dello spazio d'azione, ma non garantisce nulla sulla traiettoria prodotta dalla sequenza di token: otto token validi individualmente possono produrre una traiettoria fisicamente impossibile per collisioni, singolarità cinematiche, o violazione dei limiti di forza.

Il *secondo livello* è la fattibilità cinematica: ogni configurazione $q \in \mathbb{R}^n$ (dove n è il numero di gradi di libertà del robot) raggiunta dall'esecuzione della traiettoria deve essere interna al workspace cinematico del robot e priva di singolarità cinematiche. Un robot manipolatore ha un workspace $W = \{x \in \mathbb{R}^3 : \exists q \in Q_adm, FK(q) = x\}$ dove FK è la cinematica diretta e Q_adm è lo spazio delle configurazioni ammissibili. La fattibilità cinematica richiede che ogni posizione dell'end-effector nella traiettoria pianificata sia in W , e che la jacobiana $J(q) = \partial FK / \partial q$ abbia rango pieno lungo tutta la traiettoria — l'assenza di singolarità cinematiche che renderebbe ill-posed il problema di controllo.

Il *terzo livello*, il più stringente, è la fattibilità dinamica: la traiettoria deve essere eseguibile dalla dinamica del robot con le forze e le coppie entro i limiti attuativi. Per un robot con n gradi di libertà, la dinamica dell'equazione del moto di Lagrange è:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) = \tau$$

dove $M(q) \in \mathbb{R}^{n \times n}$ è la matrice di inerzia, $C(q, \dot{q}) \in \mathbb{R}^n$ il termine di Coriolis e centrifugo, $G(q) \in \mathbb{R}^n$ il termine gravitazionale, e $\tau \in \mathbb{R}^n$ il vettore delle coppie degli attuatori. La fattibilità dinamica richiede che $\tau \in T_adm = \{\tau : \|\tau_{il}\| \leq \tau_{max,i} \forall i\}$, dove $\tau_{max,i}$ è il limite di coppia del giunto i , lungo tutta la traiettoria — una condizione che dipende sia dalla geometria della traiettoria $(q, \dot{q}, \ddot{q})$ sia dai parametri fisici del robot (M, C, G) e del carico trasportato.

Nessuno dei meccanismi di vincolo attualmente implementati nei sistemi VLA garantisce la fattibilità ai livelli secondo e terzo. RT-2 e OpenVLA producono traiettorie che possono attraversare regioni del workspace dove il robot è in singolarità o violare i limiti di coppia degli attuatori, con conseguente fallimento dell'esecuzione o comportamento imprevedibile del controllore di basso livello.

La formulazione del problema di vincolo come MDP vincolato

La strategia naturale per imporre vincoli fisici sulla pianificazione di sistemi VLA è la formulazione come *Constrained Markov Decision Process* (CMDP), che estende l'MDP standard aggiungendo un insieme di funzioni di costo ausiliarie che devono rimanere entro soglie predefinite.

Un CMDP è una tupla $M_c = (S, A, P, r, \{c_k\}, \{d_k\}, \gamma)$ dove S è lo spazio degli stati, A lo spazio delle azioni, $P : S \times A \rightarrow \Delta(S)$ la funzione di transizione, $r : S \times A \rightarrow \mathbb{R}$ la funzione di ricompensa, $\{c_k\}_{k=1}^K$ le funzioni di costo dei vincoli, $\{d_k\}_{k=1}^K$ le soglie di costo, e γ il fattore di sconto. La policy ottimale vincolata è:

$$\pi^*_c = \operatorname{argmax}_{\pi} J_r(\pi) \text{ s.t. } J\{c_k\}(\pi) \leq d_k \quad \forall k$$

dove $J_r(\pi) = \mathbb{E}_{\pi}[\sum_t \gamma^t r(s_t, a_t)]$ è il return atteso e $J\{c_k\}(\pi) = \mathbb{E}_{\pi}[\sum_t \gamma^t c_k(s_t, a_t)]$ è il costo cumulativo del vincolo k .

Per i sistemi VLA con spazio d'azione tokenizzato, i vincoli fisici si traducono nelle seguenti funzioni di costo ausiliari:

$$\text{Vincolo di workspace: } c_{ws}(o, a) = \mathbb{1}[\text{FK}(q_t + \Delta q(a)) \notin W]$$

$$\text{Vincolo di singolarità cinematica: } c_{sing}(o, a) = \mathbb{1}[\sigma_{\min}(J(q_t + \Delta q(a))) < \epsilon_{sing}]$$

$$\text{Vincolo di forza attuativa: } c_{force}(o, a) = \max_i \max(0, \|\tau_i(q, \Delta q(a))\| - \tau_{\max,i}) / \tau_{\max,i}$$

$$\text{Vincolo di collisione: } c_{coll}(o, a) = \mathbb{1}[\exists t' > t : \|\text{FK}(q_{t'}(a)) - p_{obs}\| < r_{safe}]$$

dove $\Delta q(a)$ è la variazione di configurazione corrispondente all'azione tokenizzata a , $\text{FK}(\cdot)$ è la cinematica diretta, $J(\cdot)$ la jacobiana, $\sigma_{\min}(\cdot)$ il valore singolare minimo, p_{obs} la posizione degli ostacoli, r_{safe} il raggio di sicurezza, e $\tau_i(\cdot)$ la coppia al giunto i calcolata attraverso la cinematica inversa e la dinamica del robot.

Il problema CMDP risultante è risolvibile in linea di principio attraverso gli algoritmi di Constrained Policy Optimization (CPO) di Achiam et al. (2017) o attraverso il Lagrangian Relaxation che trasforma il problema vincolato in un problema di ottimizzazione non vincolata sul Lagrangiano aumentato:

$$L(\pi, \lambda) = J_r(\pi) - \sum_k \lambda_k (J_{\{c_k\}}(\pi) - d_k)$$

dove $\lambda_k \geq 0$ sono i moltiplicatori di Lagrange aggiornati come: $\lambda_k \leftarrow \max(0, \lambda_k + \alpha_\lambda (J_{\{c_k\}}(\pi) - d_k))$

con passo di aggiornamento α_λ . A convergenza, i moltiplicatori ottimali λ_k *quantificano il costo duale di ogni vincolo: quanto la performance sulla ricompensa degrada per ogni unità di violazione del vincolo k. Questo costo duale ha un'interpretazione fisica precisa: λ_k è il prezzo in termini di successo del task che il sistema è disposto a pagare per avvicinarsi al confine del vincolo k.*

Integrazione con gli spazi latenti JEPA: vincoli nel dominio della rappresentazione

Il framework CMDP formulato nel dominio delle azioni primitive ha un limite pratico significativo: richiede di computare in tempo reale la cinematica diretta $FK(q)$, la jacobiana $J(q)$, e la dinamica $M(q)$, $C(q, \dot{q})$, $G(q)$ — operazioni che per robot complessi con molti gradi di libertà possono richiedere decine di millisecondi, incompatibili con le frequenze di controllo di 6-10 Hz richieste dai sistemi VLA.

La strategia alternativa, coerente con il paradigma JEPA, è di riformulare i vincoli fisici direttamente nel dominio dello spazio latente dell'encoder. Se l'encoder JEPA ha appreso rappresentazioni che codificano linearmente le variabili fisiche rilevanti — posizione dell'end-effector, configurazione dei giunti, margine rispetto alle singolarità — come suggerito dai risultati di probing causale di LeWorldModel ($r^2 > 0.97$ per la posizione del blocco, Capitolo 18), allora i vincoli fisici possono essere approssimati da vincoli lineari o quadratici nello spazio latente:

$$\begin{aligned} c_{ws}(z) &\approx w_{ws}^T z + b_{ws} \geq 0 \text{ [vincolo di workspace lineare]} \\ c_{sing}(z) &\approx z^T Q_{sing} z + l_{sing}^T z \geq 0 \text{ [vincolo di singolarità quadratico]} \\ c_{force}(z, a) &\approx \|W_{force} z + w_a a - d_{force}\| \leq \tau_{bound} \text{ [vincolo di forza lineare]} \end{aligned}$$

dove i parametri (w_{ws}, b_{ws}) , (Q_{sing}, l_{sing}) , $(W_{force}, w_a, d_{force})$ sono appresi attraverso un probing lineare e quadratico sulle rappresentazioni latenti del sistema JEPA, usando come target le variabili fisiche calcolate offline dalla cinematica e dalla dinamica del robot sulle traiettorie del dataset di training.

Il costo computazionale di questo approccio è $O(d)$ per vincolo per step (dove d è la dimensionalità dello spazio latente), contro $O(n^3)$ per la cinematica e la dinamica esplicite (dove n è il numero di DOF del robot). Per $d = 512$ e $n = 7$ (Franka Panda), il guadagno computazionale è modesto in termini assoluti ma diventa significativo quando si tratta di valutare i vincoli su centinaia di traiettorie campionate durante il CEM o MPPI del Capitolo 10: valutare 500 traiettorie candidate richiede $500 \times H$ valutazioni dei vincoli ad ogni step MPC, e la differenza tra $O(d)$ e $O(n^3)$ per valutazione si moltiplica per questo fattore.

La qualità dell'approssimazione lineare/quadratica dei vincoli fisici nello spazio latente dipende criticamente da quanto linearmente le variabili fisiche siano codificate nell'encoder JEPA. I risultati empirici di LeWorldModel suggeriscono che la posizione dell'end-effector e l'orientamento del blocco siano altamente linearmente accessibili ($r^2 > 0.95$), ma non vi sono dati analoghi per le configurazioni di giunto, le coppie degli attuatori, o i margini rispetto alle singolarità — variabili più difficilmente visibili nelle osservazioni RGB e quindi potenzialmente meno lineari nello spazio latente di un encoder visivo. Il probing causale esteso su queste variabili è quindi un prerequisito necessario prima di implementare vincoli fisici nel dominio latente JEPA.

Il problema del constraint propagation nella pianificazione a catena

Un aspetto del constraints satisfaction problem che non è affrontato dalla semplice proiezione sul vocabolario valido né dalla formulazione CMDP riguarda la *propagazione dei vincoli lungo l'orizzonte di pianificazione*: un'azione valida al passo t può rendere impossibile il soddisfacimento dei vincoli al passo $t+k$, anche se l'azione era localmente fattibile. Questo è il problema del *constraint propagation* nella pianificazione a catena, analogo al backtracking nei problemi CSP combinatori ma in uno spazio continuo.

Il fenomeno si manifesta concretamente nei task di manipolazione che richiedono configurazioni cinematiche particolari. Un sistema che pianifica senza lookahead sui vincoli può scegliere una traiettoria di approccio che porta il robot in una configurazione di singolarità cinematica a metà del task, rendendo impossibile il completamento del movimento successivo anche se il punto di arrivo era all'interno del workspace. Il problema non è rilevabile dall'analisi dello stato locale ma richiede la valutazione prospettica della fattibilità dell'intero piano fino al goal.

Formalmente, si definisce la *k-step feasibility* di uno stato $(q, \dot{q})$ rispetto a un task goal g come il valore:

$$F_k(q, \dot{q}, g) = \max_{\{a_{t:t+k}\}} \min_{\{j=0\}^k} \mathbb{1}[c_{\text{phys}}(q_{t+j}, \dot{q}_{t+j}, a_{t+j}) \leq d_{\text{phys}}]$$

dove la massimizzazione è su tutte le sequenze di azioni e la minimizzazione richiede la soddisfazione simultanea dei vincoli fisici a ogni step. Se $F_k = 1$, esiste almeno un percorso fattibile di k step verso il goal; se $F_k = 0$, lo stato corrente è una *trappola cinematica* — uno stato da cui il goal non è raggiungibile entro l'orizzonte k con azioni fisicamente valide.

Il calcolo di F_k in forma esatta è computazionalmente intrattabile per $k > 2$ e per spazi d'azione continui, perché richiede l'ottimizzazione su uno spazio combinatorio esponenziale in k . Ma un'approssimazione utile per la

pianificazione è la *k-step feasibility approssimata* calcolata attraverso il World Model JEPA:

$$F_k(z_t, \omega) \approx 1 - \max_{j=0}^k P(c_{\text{phys}}(\hat{z}_{t+j}(\omega)) > d_{\text{phys}} \mid z_t)$$

dove $\hat{z}_{t+j}(\omega)$ è la traiettoria latente predetta dal World Model date le azioni dell'opzione ω , e la probabilità è calcolata attraverso la distribuzione gaussiana delle previsioni del World Model variazionale:

$$P(c_{\text{phys}}(\hat{z}_{t+j}) > d_{\text{phys}} \mid z_t) \approx \Phi((d_{\text{phys}} - \mathbb{E}[c_{\text{phys}}(\hat{z}_{t+j})]) / \sigma_{\text{phys}}(z_t, j))$$

dove Φ è la CDF gaussiana standard, $\mathbb{E}[c_{\text{phys}}]$ è il valore atteso del costo fisico sulla distribuzione predittiva del World Model, e σ_{phys} è la deviazione standard propagata attraverso la linearizzazione locale del vincolo fisico attorno alla traiettoria prevista.

Questa approssimazione converte il problema del constraint propagation da un problema di ottimizzazione combinatoria in un problema di valutazione probabilistica che può essere integrato nella funzione di costo del CEM o MPPI del Capitolo 10 come termine aggiuntivo:

$$c_{\text{total}}(\hat{z}_{t:t+H}, a_{t:t+H}) = c_{\text{task}}(\hat{z}_{t+H}) + \lambda_{\text{feas}} \cdot \sum_{j=0}^H P(c_{\text{phys}}(\hat{z}_{t+j}) > d_{\text{phys}} \mid z_t)$$

dove il secondo termine penalizza le traiettorie candidate che attraversano regioni dello spazio latente con alta probabilità di violazione dei vincoli fisici.

Il legame con i World Models gerarchici e il planning a livelli multipli

Il problema del constraint propagation ha un'interazione non triviale con la struttura gerarchica della pianificazione discussa nel Capitolo 9. In un sistema gerarchico a due livelli, il livello alto pianifica sotto-obiettivi ζ_{target} nello spazio latente astratto senza necessariamente valutare la fattibilità cinematica locale di ciascun sotto-obiettivo. Il livello basso tenta di raggiungere ζ_{target} attraverso azioni primitive, ma può scoprire che il sotto-obiettivo è cinematicamente irraggiungibile dallo stato corrente per vincoli locali che non erano visibili dal punto di vista del livello alto.

Il meccanismo di *feasibility checking inter-livello* proposto nella Sezione 9.3 — la verifica online se un sotto-obiettivo ζ_{target} è raggiungibile entro l'orizzonte H_{l-1} — è la risposta naturale a questo problema. Ma l'implementazione richiede di calcolare $F_{H_{l-1}}(z_t, \zeta_{\text{target}})$ in tempo reale, il che è computazionalmente costoso se fatto attraverso rollout completi del World Model di livello basso. Un'approssimazione praticabile è il *feasibility classifier*: una rete neurale leggera $f_{\text{feas}} : \mathbb{Z} \times \mathbb{Z} \rightarrow [0, 1]$ addestrata offline a predire la

fattibilità di un sotto-obiettivo ζ a partire dallo stato corrente z in un numero fisso di step:

$$f_feas(z_t, \zeta_{{target}}) \approx P(\exists a_{{t:t+H}} : \hat{z}_{{t+H}} \approx \zeta_{{target}} \wedge c_phys \leq d_phys \ \forall t' \in [t, t+H])$$

addestrata su un dataset di coppie $(z_t, \zeta_{{target}}, label)$ dove $label = 1$ se il sotto-obiettivo è raggiungibile con azioni fisicamente valide e $label = 0$ altrimenti, generato offline attraverso il World Model e il verificatore di vincoli. Questo feasibility classifier può essere valutato in $O(d^2)$ operazioni (forward pass di una rete MLP leggera sullo spazio latente), molto più rapidamente del rollout completo, al costo di una certa probabilità di false positive (sotto-obiettivi classificati come fattibili ma non raggiungibili) e false negative (sotto-obiettivi fattibili classificati come irraggiungibili).

Il trade-off tra precision e recall del feasibility classifier ha implicazioni dirette sulla qualità della pianificazione gerarchica: un classificatore con alta precision ma bassa recall porta il livello alto a scartare sotto-obiettivi fattibili, riducendo la qualità della pianificazione; un classificatore con alta recall ma bassa precision porta il livello basso a tentare sotto-obiettivi irraggiungibili, sprecando tempo di esecuzione e potenzialmente causando violazioni fisiche. La calibrazione del threshold del classificatore è quindi un parametro critico del sistema gerarchico che deve essere ottimizzato sul dataset di deployment.

La connessione con il scalable oversight fisico: verifica computazionale dei vincoli

Il problema del constraints satisfaction nell'action space fisico ha una risonanza inaspettata con il problema del scalable oversight discusso nel Capitolo 16. Proprio come la supervisione umana sugli output di un sistema AMI è limitata dalla competenza dell'annotatore, la verifica dei vincoli fisici di una traiettoria robotica è limitata dalla competenza del verificatore fisico — che in questo caso non è un essere umano ma il modello di cinematica e dinamica del robot.

Se il modello fisico del robot (cinematica FK, jacobiana J, dinamica M, C, G) è impreciso — il che è inevitabile per sistemi fisici reali con attrito non lineare, deformabilità dei link, incertezza nei parametri inerziali — allora i vincoli fisici computati attraverso questo modello sono anch'essi imprecisi. Un vincolo che appare soddisfatto nel modello può essere violato nel robot reale, e viceversa. Il gap di radicamento sensomotorio del Capitolo 7 è, da questo punto di vista, anche un gap di verifica dei vincoli: la verifica computazionale offline è limitata dalla qualità del modello fisico, esattamente come la supervisione umana è limitata dalla competenza degli annotatori.

La risposta strutturale a questo problema è analoga a quella del scalable oversight: stratificare la verifica per separare le violazioni che il modello può rilevare affidabilmente da quelle che richiedono verifica sensoriale online. Le

violazioni di workspace (il robot supera i limiti di articolazione) sono rilevabili con alta affidabilità dai modelli cinematici che dipendono soltanto dalla geometria del robot, invariante alle incertezze fisiche. Le violazioni di coppia agli attuatori dipendono dal carico trasportato e dall'attrito dei giunti, che variano in modo difficilmente modellabile — la loro verifica richiede sensori di coppia in linea più che calcolo offline.

Questa stratificazione porta naturalmente a una architettura di pianificazione con verifica a due livelli: un livello offline che verifica i vincoli geometrici attraverso il modello cinematico (cheap and reliable), e un livello online che monitora i vincoli dinamici attraverso i sensori del robot durante l'esecuzione (costoso ma necessario per le violazioni modello-dipendenti). Il sistema VLA, nella fase di pianificazione, ottimizza rispetto ai vincoli verificabili offline; il controllore di basso livello, durante l'esecuzione, interrompe e richiede re-pianificazione quando i sensori rilevano violazioni dei vincoli dinamici che il modello non aveva anticipato.

Questa architettura non è ancora stata implementata sistematicamente nei sistemi VLA contemporanei — RT-2 e OpenVLA eseguono le traiettorie senza monitoraggio online dei vincoli fisici — e costituisce una direzione aperta e prioritaria per il programma di ricerca futuro, con connessioni dirette sia al problema del sim-to-real gap (Sezione 8.3) sia al problema della pianificazione gerarchica sicura (Sezione 9.3) sia al problema del scalable oversight del Capitolo 16.

Il Capitolo 8 ha costruito, con la sistematicità di chi mappa un terreno accidentato piuttosto che di chi lo attraversa, la struttura matematica del sim-to-real gap: la divergenza di Jensen-Shannon tra le distribuzioni delle osservazioni simulate e reali, la sua decomposizione attraverso la cPCA contrastiva nelle direzioni dello spazio percettivo lungo cui la simulazione sottorappresenta la variabilità del mondo fisico, e la catena attraverso cui tale discrepanza si propaga dall'encoder visivo al World Model e dalla politica alle traiettorie eseguite nel deployment. Il risultato più rilevante non è il catalogo delle cause del gap — che la letteratura aveva già identificato in larga misura — ma la caratterizzazione formale del meccanismo attraverso cui la struttura percettiva dell'encoder mitiga o amplifica la propagazione: la norma di Frobenius dello Jacobiano dell'encoder, un indice misurabile attraverso differenziazione automatica, come predittore della performance degradation causata dal gap.

Ma il Capitolo 8 ha anche identificato un limite: le strategie di riduzione del gap — domain randomization, adaptation, fine-tuning con pochi campioni reali — agiscono tutte a livello della distribuzione osservativa, rendendo la distribuzione simulata più simile alla distribuzione reale oppure adattando le rappresentazioni dopo il training. Nessuna di queste strategie affronta la questione che è rimasta sullo sfondo dei Capitoli da 5 a 8: come si pianifica a lungo orizzonte in un ambiente parzialmente osservabile e fisicamente complesso, dove la policy

end-to-end dei sistemi VLA attuali è strutturalmente myope — decide l'azione corrente senza lookahead — e dove il comportamento ottimale richiede di ragionare su sequenze di azioni la cui lunghezza supera di ordini di grandezza il contesto temporale a disposizione dell'encoder? La risposta non può essere soltanto percettiva. Richiede un framework esplicito per la pianificazione come processo, non soltanto come selezione istantanea dell'azione. Richiede, in altri termini, un World Model gerarchico che operi sui fattori latenti prodotti dall'encoder JEPA anziché direttamente sulle osservazioni grezze, e che permetta all'agente di immaginare traiettorie future prima di agire. Il Capitolo 9 costruisce questo framework, partendo dalla sua formalizzazione come problema di inferenza in un modello generativo strutturato gerarchicamente.

Capitolo 9 — La pianificazione come inferenza in un modello generativo strutturato gerarchicamente

9.1 Il problema della pianificazione a lungo orizzonte: complessità e struttura

La pianificazione a lungo orizzonte è il problema di determinare una sequenza di azioni $a_0, a_1, \dots, a_{T-1}$ che porta il sistema dallo stato iniziale s_0 a uno stato goal s_G , ottimizzando una funzione cumulativa di costo lungo la traiettoria. La difficoltà computazionale di questo problema cresce esponenzialmente con l'orizzonte T e con la dimensionalità dello spazio delle azioni, rendendo la ricerca esaustiva intattrabile per qualsiasi task di complessità reale.

La letteratura ha risposto a questa difficoltà attraverso due tradizioni complementari. La tradizione del reinforcement learning ha sviluppato metodi di ottimizzazione model-free — policy gradient, Q-learning, actor-critic — che apprendono direttamente una policy $\pi : s \rightarrow a$ senza costruire esplicitamente un modello delle dinamiche. La tradizione del model-based planning ha sviluppato metodi che apprendono prima un modello delle dinamiche $f : (s, a) \rightarrow s'$ e poi pianificano attraverso questo modello usando algoritmi di ricerca o ottimizzazione. I sistemi VLA attuali — RT-2, OpenVLA — appartengono alla prima tradizione: apprendono una policy end-to-end senza un modello esplicito delle dinamiche.

La limitazione strutturale della pianificazione puramente basata su policy è la mancanza di *lookahead*: il sistema decide l'azione a_t basandosi soltanto sull'osservazione corrente o_t , senza poter «immaginare» le conseguenze di azioni alternative future. Questo non è un problema quando il task è semplice e l'orizzonte è breve — afferrare un singolo oggetto in una posizione nota richiede poca pianificazione anticipatoria — ma diventa un ostacolo

insormontabile per task con orizzonte lungo, reward sparso, e multiple sotto-sequenze di azioni che devono essere coordinate.

9.2 La pianificazione come inferenza e il ruolo del World Model

La formulazione della pianificazione come inferenza probabilistica, sviluppata da Attias (2003) e formalizzata da Toussaint (2009) e Levine (2018), offre un quadro unificato che connette pianificazione e apprendimento delle rappresentazioni in modo naturale. L'idea centrale è di definire un modello generativo $p(\tau, O \mid M)$ dove $\tau = (s_0, a_0, s_1, a_1, \dots, s_T)$ è una traiettoria, $O = (O_1, \dots, O_T)$ è una sequenza di variabili ottimali binarie, e M è il modello del mondo. La variabile ottimale $O_t = 1$ indica che l'azione a_{t-1} era ottimale allo step t — ovvero che portava verso il goal.

La distribuzione congiunta è:

$$p(\tau, O \mid M) = p(s_0) \cdot \prod_{t=0}^{T-1} [p(s_{t+1} \mid s_t, a_t; M) \cdot p(a_t \mid s_t) \cdot p(O_{t+1} = 1 \mid s_{t+1})]$$

dove $p(O_{t+1} = 1 \mid s_{t+1}) = \exp(r(s_{t+1})) / Z$ è la probabilità che lo stato s_{t+1} sia ottimale, con $r(s)$ la funzione di ricompensa. La pianificazione è allora il problema di inferenza posteriore: trovare la distribuzione $p(\tau \mid O = 1, M)$ — la distribuzione sulle traiettorie condizionata sull'evento che tutte le azioni siano state ottimali. La policy ottimale è il margine di questa distribuzione: $\pi^*(a_t \mid s_t) = \int p(\tau \mid O = 1, M) da_{0:t-1} ds_{1:t}$.

9.2.1 La variational inference come alternativa all'MCMC per la pianificazione

Il calcolo della policy ottimale attraverso l'inferenza posteriore $p(\tau \mid O = 1, M)$ è in generale computazionalmente intrattabile per modelli non lineari, a causa dell'integrale su tutti i possibili cammini τ . Due strategie di approssimazione sono state sviluppate nella letteratura: l'approssimazione variazionale e l'approssimazione Monte Carlo. Il loro confronto illumina quale sia più naturale nel contesto dei World Models basati su JEPA.

L'approssimazione variazionale introduce una distribuzione ausiliaria $q(\tau)$ e ottimizza il bound inferiore:

$$\begin{aligned} \log p(O = 1 \mid M) &\geq \mathbb{E}_{q(\tau)} [\log p(O = 1, \tau \mid M)] - \mathbb{E}_{q(\tau)} [\log q(\tau)] \\ &= \mathbb{E}_{q(\tau)} [\sum_t r(s_{t+1})] - D_{KL}(q(\tau) \parallel p(\tau \mid M)) \end{aligned}$$

dove il secondo termine è la divergenza KL tra la distribuzione variazionale sulle traiettorie e la distribuzione a priori sul modello di mondo. Questa è esattamente la struttura dell'obiettivo RL con regolarizzazione entropica

(MaxEnt RL), e la policy variazionale ottimale è la distribuzione di Boltzmann sul return:

$$q^*(a_t | s_t) \propto p(a_t | s_t) \cdot \exp(Q^*(s_t, a_t) / \beta)$$

dove $Q^*(s_t, a_t)$ è il Q-value ottimale e β è la temperatura. Questa soluzione mostra che la pianificazione come inferenza variazionale è equivalente al soft Q-learning con temperatura β — un collegamento formale tra pianificazione probabilistica e RL che chiarisce perché i metodi di model-based RL basati su World Models raggiungono comportamenti qualitativamente simili ai metodi model-free pur operando su principi diversi.

Nel contesto JEPA, il World Model M opera nello spazio latente: M è la distribuzione $p(\hat{z}_{t+1} | \hat{z}_t, a_t)$ appresa dal predittore. La distribuzione variazionale $q(\tau)$ diventa una distribuzione su traiettorie latenti $\{\hat{z}_0, a_0, \hat{z}_1, a_1, \dots\}$, e la pianificazione variazionale ottimizza il return latente $\mathbb{E}[\sum_t r(\hat{z}_{t+1})]$ dove $r(\hat{z})$ è una funzione di costo definita nello spazio latente. Questo framework si riduce al ciclo MPC derivato nel Capitolo 10 quando la distribuzione variazionale è concentrata su una singola traiettoria ottimale — il caso deterministico del soft planning.

Il World Model M — il modello della funzione di transizione $p(s_{t+1} | s_t, a_t; M)$ — è il componente cruciale di questa formulazione. Senza un World Model, l'inferenza posteriore richiede di interagire con l'ambiente reale per valutare le conseguenze delle azioni alternative. Con un World Model accurato, la pianificazione può avvenire interamente «nella testa» del sistema, attraverso simulazione mentale delle traiettorie alternative, senza nessuna interazione con l'ambiente fisico.

9.2.2 La derivazione formale di MPPI come inferenza variazionale con temperatura di softmax

Il *Model Predictive Path Integral* (MPPI), proposto da Williams et al. (2016) e formalizzato nella sua versione con temperatura nel successivo lavoro del 2018, costituisce l'istanza più rigorosa e computazionalmente trattabile della pianificazione come inferenza variazionale nel dominio del controllo continuo. La sua derivazione a partire dalla struttura variazionale delineata nella sezione 9.2.1 è istruttiva perché mostra in modo esplicito come la scelta della distribuzione di riferimento — solitamente ignorata nelle presentazioni algoritmiche di MPPI — determini tanto le proprietà di convergenza quanto il significato fisico della temperatura λ .

Sia $\tau = (x_0, u_0, x_1, u_1, \dots, x_{T-1}, u_{T-1}, x_T)$ una traiettoria nello spazio degli stati e delle azioni. Il modello delle dinamiche è $x_{t+1} = f(x_t, u_t) + \epsilon_t$, dove $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ rappresenta il rumore stocastico del sistema. La distribuzione a priori sulle sequenze di controllo è la gaussiana di prodotto $p(U) = \prod_{t=0}^{T-1} \mathcal{N}(u_t; \bar{u}_t, \Sigma_u)$, con $\bar{u}_t$ la sequenza nominale corrente e

Σ_u la matrice di covarianza del processo di controllo. Definita la variabile di costo:

$$S(\tau) = \Phi(x_T) + \sum_{t=0}^{T-1} q(x_t, u_t)$$

dove Φ è il costo terminale e q è il costo di stato-azione a ogni passo, la densità di importanza ottimale che massimizza il bound variazionale della sezione 9.2.1 è, come si mostrerà, la distribuzione di Boltzmann su U condizionata alle dinamiche:

$$q^*(U) \propto p(U) \cdot \exp(-S(\tau(U)) / \lambda)$$

La derivazione procede in tre passi. Nel primo, si scrive il bound variazionale applicato al modello generativo $p(\tau, O = 1 \mid M)$ della sezione 9.2:

$$\log p(O = 1 \mid M) \geq \mathbb{E}_{\tau} \{q(\tau)\} [\log p(O = 1, \tau \mid M)] - \mathbb{E}_{\tau} \{q(\tau)\} [\log q(\tau)] = \mathbb{E}_{\tau} \{q(\tau)\} [\log p(O = 1 \mid \tau)] - D_{\text{KL}}(q(\tau) \parallel p(\tau \mid M))$$

Usando $p(O_{t+1} = 1 \mid x_{t+1}) = \exp(-q(x_t, u_t) / \lambda) \cdot \exp(-\Phi(x_T) / \lambda) / Z$ (con Z la costante di normalizzazione dipendente solo da λ) e sostituendo la distribuzione delle dinamiche deterministiche condizionata sulla sequenza di controllo U — il che trasforma l'integrale su τ in un integrale su U — il bound diventa:

$$\mathcal{L}(q) = \mathbb{E}_U \{q(U)\} [-S(\tau(U)) / \lambda] - D_{\text{KL}}(q(U) \parallel p(U)) + \text{costanti}$$

Nel secondo passo, si massimizza $\mathcal{L}(q)$ rispetto a $q(U)$. Per la struttura generale del problema di minimizzazione della KL con prior $p(U)$ e ricompensa $F(U) = -S(\tau(U)) / \lambda$, il massimizzatore analitico è la distribuzione esponenziale:

$$q^*(U) = p(U) \cdot \exp(-S(\tau(U)) / \lambda) / Z_{\lambda}$$

dove $Z_{\lambda} = \int p(U) \cdot \exp(-S(\tau(U)) / \lambda) dU$ è il fattore di normalizzazione. Questa è esattamente la densità di importance sampling ottimale di MPPI. Il parametro λ che compare nella temperatura dell'esponenziale è il medesimo λ che compare nel regolarizzatore entropico del MaxEnt RL — non è una coincidenza ma un'identità formale: entrambi derivano dallo stesso principio di dualità tra ottimizzazione dell'obiettivo atteso e penalizzazione della deviazione dal prior.

Nel terzo passo, si derivano le statistiche sufficienti di $q^*(U)$. Poiché il prior $p(U)$ è gaussiano con media $\bar{u}$ e covarianza Σ_u , e la distribuzione ottimale è la prodotto del prior per un peso, la media ottimale è:

$$\mathbb{E}_U \{q^*\} [u_t] = \bar{u}_t + \Sigma_u \cdot \mathbb{E}_U \{q^*\} [\nabla \{u_t\} \log q^*(U) - \nabla \{u_t\} \log p(U)]$$

che si semplifica, usando il trucco dell'identità del score per distribuzioni esponenziali, a:

$$\mathbb{E}_{\{q^*\}}[u_t] = \bar{u}_t + \mathbb{E}_{\{q^*\}} [\delta u_t \cdot (-S(\tau) / \lambda)] / Z_\lambda$$

dove $\delta u_t = u_t - \bar{u}_t$ è la perturbazione rispetto alla sequenza nominale. Questa è la formula di aggiornamento di MPPI per la media dei controlli, che in forma campionata (con K traiettorie campionate dal prior $p(U)$) diventa l'algoritmo:

$$\bar{u}_t \leftarrow \bar{u}_t + \sum_{k=1}^K w_k \cdot \delta u_t^{\{(k)\}}$$

dove i pesi normalizzati sono:

$$w_k = \exp(-S(\tau^{\{(k)\}}) / \lambda) / \sum_{j=1}^K \exp(-S(\tau^{\{(j)\}}) / \lambda)$$

ed è immediatamente riconoscibile la struttura della *softmax* con temperatura λ applicata ai costi delle traiettorie campionate. La temperatura λ controlla il grado di concentrazione dei pesi: per $\lambda \rightarrow \infty$ i pesi tendono all'uniforme $w_k = 1/K$ e l'aggiornamento è la media aritmetica semplice delle perturbazioni, ignorando completamente la struttura dei costi; per $\lambda \rightarrow 0$, come si mostrerà nella sezione 9.2.3, i pesi si concentrano sulla traiettoria di costo minimo — il limite deterministico del Cross-Entropy Method.

Il ruolo della temperatura λ merita un'analisi che va al di là della semplice formulazione algoritmica. Si consideri la decomposizione dell'obiettivo variazionale $\mathcal{L}(q)$:

$$\mathcal{L}(q) = -(1/\lambda) \cdot \mathbb{E}_{\{q(U)\}} [S(\tau(U))] + H(q(U)) - D_{\text{KL}}(q(U) \parallel p(U)) + \text{costanti}$$

dove $H(q) = -\mathbb{E}_q[\log q]$ è l'entropia della distribuzione variazionale. Il termine $-D_{\text{KL}}(q \parallel p) = H(q) - \mathbb{E}_q[\log p]$ incorpora sia il termine entropico sia il log-prior, cosicché l'obiettivo totale è:

$$\mathcal{L}(q) = -(1/\lambda) \cdot \mathbb{E}_q[S(\tau)] + \mathbb{E}_q[\log p(U)] + H(q(U)) + \text{costanti}$$

Il primo termine incentiva la concentrazione di q sulle traiettorie di basso costo. Il secondo termine incentiva q a rimanere vicina al prior nominale $p(U)$. Il terzo termine entropico incentiva la dispersione di q . La temperatura λ bilancia i primi due termini rispetto al terzo: a λ piccolo, la pressione verso il costo minimo domina l'entropia e si ottiene una distribuzione concentrata; a λ grande, l'entropia domina e si ottiene esplorazione diffusa. Questa è esattamente la struttura del trade-off esplorazione-sfruttamento nel MaxEnt RL, formalizzata come equazione di bilanciamento nel sistema di pianificazione.

Il legame con il MaxEnt RL si rende ancora più esplicito quando si passa dalla pianificazione a orizzonte finito alla politica a orizzonte infinito. Nel MaxEnt RL, la funzione di valore ottimale soddisfa l'equazione di Bellman morbida:

$$V^*(x) = \text{soft max}_a \{ r(x, a) + \gamma \mathbb{E}_{\{x' | x, a\}} [V^*(x')]] = \lambda \cdot \log \int \exp([r(x, a) + \gamma \mathbb{E}_{\{x' | x, a\}} [V^*(x')]] / \lambda) da + \lambda \cdot \log Z_\pi$$

dove $\text{soft max}_{\lambda}[\cdot]$ è l'operatore di log-sum-exp con temperatura λ , che è il limite continuo dell'operatore di softmax discreto. La corrispondenza con la struttura di MPPI è completa: l'operatore di softmax nella formula di aggiornamento dei controlli di MPPI è esattamente la versione a orizzonte finito dell'operatore di Bellman morbida del MaxEnt RL. Il valore campionato di MPPI approssima il valore della politica ottimale MaxEnt RL con un estimatore Monte Carlo del log-sum-exp:

$$V_{\text{MPPI}}(x_0) \approx -\lambda \cdot \log \left(\frac{1}{K} \sum_{k=1}^K \exp(-S(\tau^{\{k\}}) / \lambda) \right)$$

che converge al valore MaxEnt esatto per $K \rightarrow \infty$ grazie alla legge dei grandi numeri per l'estimatore log-somma-esponenziale.

Questa identità ha conseguenze dirette per il contesto JEPA. Quando il World Model opera nello spazio latente — ovvero quando le dinamiche sono formulate come $x_{t+1} = f(\hat{z}_t, a_t)$ con $\hat{z}_t = E(x_t)$ la rappresentazione prodotta dall'encoder — il costo $S(\tau)$ diventa un funzionale sulle traiettorie latenti $\{\hat{z}_0, a_0, \hat{z}_1, \dots, \hat{z}_T\}$. L'equazione di aggiornamento di MPPI rimane formalmente identica, ma i pesi w_k sono calcolati sui costi delle traiettorie latenti campionate attraverso il predittore JEPA. La temperatura λ assume qui un significato epistemico aggiuntivo: essa modella l'incertezza del World Model — un World Model poco accurato dovrebbe usare λ elevato per non concentrare i pesi su traiettorie che il modello valuta erroneamente come ottimali, mentre un modello accurato può usare λ basso per sfruttare le differenze fini tra traiettorie.

9.2.3 La derivazione del CEM come caso limite deterministico di MPPI per $\lambda \rightarrow 0$

Il *Cross-Entropy Method* (CEM), applicato alla pianificazione da Botev et al. (2013) e popularizzato nel contesto dei World Models da Hafner et al. (2019) con PlaNet, si presenta superficialmente come un algoritmo distinto da MPPI: invece di pesare tutte le traiettorie con pesi morbidi, CEM seleziona le migliori in percentuale (le «elite», tipicamente il 10%–20% delle K traiettorie campionate), calcola la media e la varianza delle sequenze di controllo elite, e usa queste statistiche per aggiornare la distribuzione di campionamento. Tuttavia, questa distinzione è puramente di grado e non di natura: il CEM emerge esattamente come il caso limite $\lambda \rightarrow 0$ di MPPI, ovvero come la pianificazione variazionale nel regime di temperatura nulla, in cui la distribuzione di Boltzmann degenera in una distribuzione concentrata sulle traiettorie ottimali.

Si dimostri formalmente questo limite. I pesi MPPI sono:

$$w_k(\lambda) = \exp(-S(\tau^{\{k\}}) / \lambda) / \sum_{j=1}^K \exp(-S(\tau^{\{j\}}) / \lambda)$$

Sia $k^* = \operatorname{argmin}_k S(\tau^{\{k\}})$ l'indice della traiettoria di costo minimo tra le K campionate. La tecnica standard per calcolare il limite della softmax è sottrarre il minimo dal numeratore e dal denominatore (riscaldamento per la stabilità numerica):

$$w_k(\lambda) = \frac{\exp(-(S(\tau^{\{k\}}) - S(\tau^{\{k^*\}})) / \lambda)}{\sum_{j=1}^K \exp(-(S(\tau^{\{j\}}) - S(\tau^{\{k^*\}})) / \lambda)}$$

Per ogni $k \neq k^*$, il termine $(S(\tau^{\{k\}}) - S(\tau^{\{k^*\}})) > 0$ (assumendo che il minimo sia unico), cosicché il numeratore $\exp(-(S(\tau^{\{k\}}) - S(\tau^{\{k^*\}})) / \lambda) \rightarrow 0$ esponenzialmente velocemente per $\lambda \rightarrow 0$. Per $k = k^*$, il numeratore vale 1 per qualsiasi λ . Nel limite:

$$\lim_{\lambda \rightarrow 0} w_k(\lambda) = \delta_{\{k, k^*\}}$$

ovvero i pesi degenerano nel vettore canonico di Kronecker: tutto il peso si concentra sull'unica traiettoria di costo minimo. L'aggiornamento di MPPI nel limite diventa:

$$\bar{u}_t \leftarrow \bar{u}_t + \delta u_t^{\{k^*\}} = u_t^{\{k^*\}}$$

il che significa che la sequenza di controllo nominale viene rimpiazzata interamente dalla sequenza di controllo della traiettoria ottimale campionata — non una media pesata, ma la selezione diretta del vincitore.

Questa concentrazione sul singolo vincitore è il comportamento qualitativo del CEM, ma la corrispondenza è ancora più precisa. Il CEM standard seleziona le migliori K_e traiettorie (con $K_e = \lfloor \varrho \cdot K \rfloor$ per una frazione elite $\varrho \in (0,1)$), non la sola traiettoria ottimale. Questo corrisponde al caso in cui il campionamento finito fa sì che più traiettorie si trovino in un intorno del minimo — un'inevitabilità statistica per K finito. Si può formalizzare il CEM come il limite $\lambda \rightarrow 0$ con raffreddamento graduale (temperature scheduling): si parte da λ elevato, si campionano K traiettorie, si aggiorna la media, si riduce λ di un fattore geometrico $\alpha \in (0,1)$, e si itera. Al limite di $\lambda \rightarrow 0$, tutte le traiettorie non-elite hanno peso $w_k \rightarrow 0$, e i pesi uniformi $w_k = 1/K_e$ delle elite emergono come la discretizzazione della distribuzione uniforme ristretta all'insieme $\{\tau^{\{k\}} : k \in \text{elite}\}$.

Un modo equivalente per vedere la stessa identità è notare che l'operatore di softmax per $\lambda \rightarrow 0$ converge all'operatore max:

$$-\lambda \cdot \log \sum_{k=1}^K \exp(-S(\tau^{\{k\}}) / \lambda) \rightarrow \min_k S(\tau^{\{k\}})$$

Questa è la dualità soft-hard tra log-sum-exp e min/max: l'operatore MPPI è la versione morbida dell'ottimizzatore deterministico, e λ è il parametro di interpolazione. Nella tabella seguente, sintetizzata informalmente per chiarezza espositiva, si organizza questa relazione:

— Per $\lambda \rightarrow \infty$: MPPI degenera nella media semplice delle perturbazioni (piena esplorazione, nessun sfruttamento del gradiente di costo). — Per λ finito e moderato: MPPI è il pianificatore variazionale MaxEnt con esplorazione bilanciata. — Per $\lambda \rightarrow 0$: MPPI converge al CEM deterministico con selezione delle elite. — Per $K = 1, \lambda \rightarrow 0$: MPPI converge al gradient descent sul costo (se il modello è differenziabile), ovvero alla pianificazione basata su backpropagation through time.

Questa struttura gerarchica di regimi è importante per il contesto JEPa perché suggerisce una strategia adattiva per la scelta di λ in funzione dell'accuratezza del World Model. Quando il predittore JEPa è appena inizializzato e le previsioni latenti $\hat{z}_{t+1} = g(\hat{z}_t, a_t)$ sono poco accurate, un λ elevato è appropriato — esso equivale a dare fiducia modesta alle differenze di costo tra traiettorie, mantenendo un'esplorazione larga. Man mano che il predittore converge e l'accuratezza delle previsioni cresce (monitorabile attraverso la perdita di predizione JEPa sulle sequenze di validazione), λ può essere ridotto progressivamente — avvicinando il pianificatore al CEM deterministico che sfrutta appieno la struttura del modello appreso. Questo schema di *temperature annealing adattivo* è la versione continua dello scheduling di temperatura nel MaxEnt RL e trova una giustificazione bayesiana precisa nel framework della pianificazione come inferenza: λ è inversamente proporzionale alla fiducia del modello, e la riduzione graduale di λ corrisponde all'aggiornamento bayesiano sulla qualità del World Model conforme si accumulano dati.

Una proprietà tecnica importante del CEM, che non è immediatamente visibile nella presentazione come caso limite di MPPI ma emerge chiaramente dalla derivazione variazionale, è la sua corrispondenza con la minimizzazione della divergenza KL *inversa* (mode-seeking) piuttosto che della KL *diretta* (mean-seeking). Nella formulazione variazionale, MPPI minimizza $D_{KL}(q \parallel p_{\text{Boltzmann}})$ — la KL dalla distribuzione variazionale q alla distribuzione obiettivo $p_{\text{Boltzmann}}$ — il che è una divergenza diretta e produce distribuzioni q che coprono i modi di $p_{\text{Boltzmann}}$ (mean-seeking). Nel limite $\lambda \rightarrow 0$ in cui $p_{\text{Boltzmann}}$ si concentra sul minimo globale, la distinzione tra KL diretta e KL inversa svanisce perché la distribuzione obiettivo è un punto di massa: entrambe producono la concentrazione su k^* . Tuttavia, per λ finito, la scelta di quale KL minimizzare ha conseguenze: MPPI, come minimizzatore della KL diretta, è in principio capace di esplorare distribuzioni multimodali e trovare alternative quasi-ottimali, mentre il CEM (come suo limite deterministico) per definizione converge verso un unico modo della distribuzione di costo. Questa distinzione diventa rilevante nei landscape di pianificazione con molteplici traiettorie quasi-equivalenti, che sono caratteristici dei task di manipolazione con ridondanza cinematica — un'osservazione con implicazioni dirette per i sistemi VLA analizzati nei capitoli precedenti.

Completata la derivazione del CEM come caso limite di MPPI, si ha dunque un quadro unificato della pianificazione campionata: un'unica famiglia di pianificatori parametrizzata dalla temperatura λ e dalla frazione elite q , in cui

MPPI con λ finito e CEM con $\lambda \rightarrow 0$ sono istanze di un'unica struttura variazionale. Nel contesto del Capitolo 10, dove il MPC con World Model JEPA viene analizzato operativamente, questa unificazione suggerisce che la scelta tra MPPI e CEM non è una scelta tra algoritmi qualitativamente diversi ma una scelta di iperparametro nel continuum della temperatura, con conseguenze documentabili sulla traiettoria di convergenza del pianificatore.

Sezione 9.3 — La struttura gerarchica: separare il tempo e l'astrazione

Il problema della pianificazione a lungo orizzonte non è superabile semplicemente estendendo l'orizzonte H del ciclo MPC descritto nel Capitolo 10. L'accumulo dell'errore di previsione nel World Model cresce con la profondità del rollout con un tasso almeno polinomiale — come dimostrato nel Capitolo 11 attraverso il bound sulla subottimalità — e i metodi di campionamento come CEM e MPPI richiedono un numero di traiettorie campionate che cresce esponenzialmente con l'orizzonte per mantenere la stessa copertura dello spazio delle azioni. Nessuna di queste difficoltà è contingente: derivano dalla struttura matematica del problema di ottimizzazione, non dall'insufficienza degli algoritmi correnti.

La risposta della letteratura è la *separazione delle scale temporali di pianificazione*, che nella tradizione del Reinforcement Learning gerarchico assume la forma dei framework basati su opzioni (Sutton et al., 1999), MAXQ (Dietterich, 2000) e Feudal Networks (Vezhnevets et al., 2017). In ciascuno di questi framework, il ciclo di controllo è organizzato in livelli che operano a frequenze diverse: i livelli bassi producono azioni primitive ad alta frequenza; i livelli intermedi selezionano sotto-obiettivi a frequenza ridotta; i livelli alti pianificano sequenze di macro-azioni su orizzonti temporali estesi. Il guadagno computazionale di questa organizzazione non è soltanto pratico ma strutturale: la scomposizione gerarchica trasforma un problema di pianificazione di lunghezza T in una catena di sotto-problemi di lunghezza T/k , il che riduce la complessità computazionale da esponenziale in T a esponenziale in T/k moltiplicata per il numero di livelli.

La formalizzazione a livelli multipli e il modello di Markov semi-deciso

Si considera un framework gerarchico a L livelli indicizzati da $l \in \{1, 2, \dots, L\}$, dove $l = 1$ è il livello delle azioni primitive e $l = L$ è il livello della pianificazione di più alto grado di astrazione. A ciascun livello l è associata una *timescale* τ_l : il livello l opera su intervalli temporali di lunghezza $\tau_l = \tau_1 \cdot k^{l-1}$, dove τ_1 è il periodo campionario delle azioni primitive e $k > 1$ è il *fattore di compressione temporale* tra livelli consecutivi. Per $k = 5$ e $\tau_1 = 50$ ms, si ha $\tau_2 = 250$ ms, $\tau_3 = 1.25$ s, $\tau_4 = 6.25$ s — scale temporali che corrispondono grosso modo a riflessi muscolari, movimenti coordinati, sotto-task elementari e task complessi,

rispecchiando la stratificazione della pianificazione motoria umana documentata dalla neuroscienza computazionale.

Il processo decisionale a ciascun livello l è formalizzato come un Processo di Markov Semi-Deciso (SMDP), la cui struttura differisce da un MDP standard per il fatto che le azioni di livello l sono *opzioni* $\omega^{\{l\}} \in \Omega^{\{l\}}$ eseguite per un numero variabile di step del livello inferiore. Un'opzione $\omega^{\{l\}} = (\mathcal{F}^{\{l\}}_{\omega}, \pi^{\{l\}}_{\omega}, \beta^{\{l\}}_{\omega})$ è specificata da tre componenti: l'insieme degli stati ammissibili $\mathcal{F}^{\{l\}}_{\omega} \subseteq Z^{\{l\}}$, la policy intra-opzione $\pi^{\{l\}}_{\omega} : Z^{\{l\}} \times A^{\{l\}} \rightarrow [0,1]$ che seleziona le azioni del livello $l-1$ durante l'esecuzione, e la funzione di terminazione $\beta^{\{l\}}_{\omega} : Z^{\{l\}} \rightarrow [0,1]$ che determina la probabilità di termine dell'opzione nello stato corrente. La sequenza temporale di esecuzione è la seguente: il livello l seleziona un'opzione $\omega^{\{l\}}_k$ a ogni istante di switch t_k ; l'opzione esegue la propria policy intra-opzione per un numero di step del livello $l-1$ dato dalla funzione di terminazione; quando $\beta^{\{l\}}_{\omega}(z^{\{l\}}) \geq u$ con $u \sim \text{Uniform}(0,1)$, l'opzione termina e il controllo torna al livello l che seleziona la prossima opzione.

La struttura del World Model gerarchico associato a questo framework ha $L-1$ componenti, uno per ogni transizione inter-livello. Il World Model al livello l modella le transizioni tra stati dello spazio latente $Z^{\{l\}}$ condizionate sull'opzione selezionata:

$$f^{\{l\}}_{\varphi} : Z^{\{l\}} \times \Omega^{\{l\}} \rightarrow Z^{\{l\}}$$

che stima il successore dello stato latente al livello l dopo l'esecuzione completa dell'opzione $\omega^{\{l\}}$ a partire dallo stato $z^{\{l\}}_k$:

$$z^{\{l\}}_{k+1} \approx f^{\{l\}}_{\varphi}(z^{\{l\}}_k, \omega^{\{l\}}_k) + \varepsilon^{\{l\}}_{k+1}$$

dove $\varepsilon^{\{l\}}_{k+1} \sim \mathcal{N}(0, \sigma^{\{l\}^2} \cdot I)$ è il rumore di previsione al livello l . Poiché ogni livello opera su una timescale $\tau_l = \tau_1 \cdot k^{\{l-1\}}$, il World Model $f^{\{l\}}_{\varphi}$ comprime l'evoluzione temporale di $k^{\{l-1\}}$ step primitivi in una singola previsione nello spazio astratto $Z^{\{l\}}$ — una forma di *temporal abstraction* che non è imposta dall'architettura ma emerge naturalmente dall'obiettivo di previsione al livello corretto di astrazione.

Il mapping tra spazi latenti JEPA a scale diverse

La connessione tra il framework gerarchico e il paradigma JEPA non è formale ma strutturalmente motivata dalla seguente osservazione. L'encoder V-JEPA, nella sua versione multi-scala, produce rappresentazioni a diverse profondità della rete: i layer iniziali catturano strutture locali ad alta risoluzione spaziale (bordi, texture, gradienti di luminanza), i layer intermedi catturano relazioni spaziali tra regioni (forme, disposizioni relative di parti), i layer profondi catturano strutture semantiche di alto livello (categorie, relazioni funzionali tra oggetti). Questa stratificazione percettiva corrisponde naturalmente alla stratificazione temporale del framework gerarchico: le rappresentazioni a bassa

risoluzione semantica sono appropriate per il controllo ad alta frequenza che opera su brevi orizzonti, mentre le rappresentazioni a elevata astrazione semantica sono appropriate per la pianificazione a lungo orizzonte che opera su scale temporali estese.

Si definisca formalmente questa corrispondenza. Siano $E^{\{1\}}_{\theta}, E^{\{2\}}_{\theta}, \dots, E^{\{L\}}_{\theta}$ le funzioni di embedding prodotte dai layer dell'encoder JEPA a profondità crescente, con $E^{\{l\}}_{\theta} : \mathcal{X} \rightarrow Z^{\{l\}} \subseteq \mathbb{R}^{d_l}$. La sequenza di dimensionalità $d_1 \geq d_2 \geq \dots \geq d_L$ riflette la progressiva compressione dell'informazione percettiva: le rappresentazioni dei layer bassi mantengono alta risoluzione spaziale (d_1 grande), mentre le rappresentazioni dei layer profondi sono più compatte (d_L piccolo, tipicamente la dimensionalità dello spazio globale dell'encoder). Nella pratica di V-JEPA 2.1, le Deep Self-Supervision layers (ai blocchi 12, 24, 36 e 48 di un ViT-G) producono esattamente questa struttura a quattro livelli, il che rende V-JEPA 2.1 un candidato naturale per il Perception Module gerarchico senza modifiche architetturali.

Lo stato latente al livello gerarchico l è ottenuto aggregando temporalmente le rappresentazioni al livello $l-1$. Si definisca l'aggregazione tramite l'operatore $AGG^{\{l\}} : (Z^{\{l-1\}})^k \rightarrow Z^{\{l\}}$ che comprime k stati latenti consecutivi del livello $l-1$ in un singolo stato latente del livello l :

$$\zeta^{\{l\}}_n = AGG^{\{l\}}(\zeta^{\{l-1\}}_{(n-1)k+1}, \zeta^{\{l-1\}}_{(n-1)k+2}, \dots, \zeta^{\{l-1\}}_{nk})$$

Un'implementazione naturale di $AGG^{\{l\}}$ è una forma ridotta del transformer aggregatore di seq-JEPA: si prendono i k embedding consecutivi come input del transformer (senza token di azione, poiché l'aggregazione temporale avviene sull'intero intervallo di esecuzione dell'opzione), e si usa il token [AGG] come output. Questa scelta preserva la struttura architetturale del paradigma JEPA — lo stesso tipo di aggregatore usato per combinare viste multiple viene riutilizzato per combinare frame temporalmente distanziati — e consente di trasferire, almeno in linea di principio, l'invarianza appresa durante il preaddestramento seq-JEPA alla compressione temporale gerarchica.

La perdita di previsione gerarchica e la sua derivazione variazionale

L'addestramento congiunto dei World Model ai diversi livelli gerarchici richiede un obiettivo che incentivi la predittività a ogni scala temporale simultaneamente. Si propone la seguente *perdita di previsione gerarchica*, derivata come lower bound variazionale sulla verosimiglianza delle traiettorie multi-livello:

$$\mathcal{L}_{\text{hier}} = \sum_{l=1}^L \alpha_l \cdot \mathcal{L}^{\{l\}}_{\text{pred}}$$

dove α_l sono i pesi di bilanciamento tra i livelli (tipicamente α_l decrescente con l , per privilegiare la precisione al livello basso), e $\mathcal{L}^{\{l\}}_{\text{pred}}$ è la perdita di previsione al livello l :

$$\mathcal{L}^{\{l\}}_{pred} = \mathbb{E} \{ (z^{\{l\}}_{-k}, \omega^{\{l\}}_k, \hat{z}^{\{l\}}_{k+1}) \sim D^{\{l\}} \mid \|\hat{f}^{\{l\}}_{-k}(z^{\{l\}}_{-k}, \omega^{\{l\}}_k) - \text{sg}(\hat{z}^{\{l\}}_{k+1})\|^2 \}$$

dove $D^{\{l\}}$ è il dataset di transizioni al livello l (coppie stato-opzione-successore estratte dalle traiettorie di training), e $\text{sg}(\cdot)$ è lo stop-gradient che previene il collasso. La presenza dello stop-gradient è essenziale come nel caso mono-livello: senza di esso, il World Model e l'encoder potrebbero colludere producendo rappresentazioni costanti che minimizzano la perdita senza apprendere nulla di utile sulla struttura dinamica del dominio.

La derivazione variazionale di questa perdita procede per analogia con la derivazione dell'ELBO di Var-JEPA. Si introduce la distribuzione variazionale $q^{\{l\}}(\hat{z}^{\{l\}}_{k+1} \mid z^{\{l\}}_{-k}, \omega^{\{l\}}_k)$ che approssima il posteriore vero $p(\hat{z}^{\{l\}}_{k+1} \mid z^{\{l\}}_{-k}, \omega^{\{l\}}_{-k})$ sulle transizioni al livello l . Il lower bound sulla log-verosimiglianza delle traiettorie al livello l è:

$$\log p(z^{\{l\}}_{1:T^{\{l\}}}) \geq \sum_{k=1}^{T^{\{l\}}-1} \left[\mathbb{E}_{q^{\{l\}}} [\log p^{\{l\}}(z^{\{l\}}_{k+1} \mid z^{\{l\}}_{-k}, \omega^{\{l\}}_{-k})] - D_{\text{KL}}(q^{\{l\}}(\hat{z}^{\{l\}}_{k+1} \mid \cdot) \parallel p^{\{l\}}_{\text{prior}}(\hat{z}^{\{l\}}_{k+1})) \right]$$

dove $p^{\{l\}}_{\text{prior}}$ è il prior gaussiano isotropo imposto da SIGReg al livello l . Quando il World Model $\hat{f}^{\{l\}}_{-k}$ è gaussiano con media $\hat{f}^{\{l\}}_{-k}(z^{\{l\}}_{-k}, \omega^{\{l\}}_k)$ e varianza $\sigma^{\{l\}^2} \cdot I$, il termine di log-likelihood si semplifica nell'errore quadratico scalato: $\log p^{\{l\}} \approx -\|\hat{f}^{\{l\}}_{-k}(z^{\{l\}}_{-k}, \omega^{\{l\}}_k) - \hat{z}^{\{l\}}_{k+1}\|^2 / (2\sigma^{\{l\}^2})$. Sommando su tutti i livelli con i pesi $\alpha_{l-1} = 1/(2\sigma^{\{l\}^2})$ si recupera esattamente la forma $\mathcal{L}_{\text{hier}}$ sopra definita.

La comunicazione inter-livello: goal conditioning e temporal consistency

Una componente cruciale del framework gerarchico che non emerge automaticamente dalla formalizzazione SMDP è il meccanismo di comunicazione tra livelli adiacenti: come il livello l comunica al livello $l-1$ il sotto-obiettivo che deve raggiungere, e come il livello $l-1$ informa il livello l dello stato del mondo al termine dell'esecuzione dell'opzione.

La strategia di *goal conditioning* naturale nel paradigma JEPA è la seguente: il livello l seleziona un sotto-obiettivo come vettore nello spazio latente $Z^{\{l-1\}}$, cioè $\zeta^{\{l-1\}}_{\text{target}} \in Z^{\{l-1\}}$, e lo trasmette al livello $l-1$ come target per la pianificazione. Il livello $l-1$, a sua volta, pianifica la sequenza di azioni primitive che minimizza la distanza nello spazio latente tra lo stato corrente e il sotto-obiettivo:

$$a^*_{t:t+H\{l-1\}} = \underset{a}{\text{argmin}} \{ a_{t:t+H\{l-1\}} \mid \|z^{\{l-1\}}_{t+H\{l-1\}} - \zeta^{\{l-1\}}_{\text{target}}\|^2 \}$$

dove $z^{\{l-1\}}_{t+H\{l-1\}}$ è lo stato latente predetto dal World Model di livello $l-1$ alla fine dell'orizzonte di pianificazione H_{l-1} . Questa formulazione è

consistente con la struttura dell'MPC latente del Capitolo 10: il ciclo predizione-ottimizzazione-esecuzione rimane identico, ma ora il costo di pianificazione è la distanza dallo spazio latente del livello $l-1$ a un target fornito dal livello superiore piuttosto che da una funzione di costo task-specifica.

La *temporal consistency* tra livelli è garantita dalla coerenza tra le rappresentazioni prodotte dall'encoder a scale diverse. Se l'encoder è addestrato con l'obiettivo di Deep Self-Supervision di V-JEPA 2.1, le rappresentazioni ai layer 12, 24, 36 e 48 sono mutuamente coerenti nel senso che informazione causale presente al layer 48 (l'uscita) è stata necessariamente preservata attraverso i layer intermedi (come dimostrato empiricamente nella riduzione del divario tra probing multi-layer e probing last-layer documentata nella Tabella 12 del paper). Questa coerenza verticale dell'encoder è il prerequisito affinché la comunicazione inter-livello non perda informazione critica nella trasmissione dal livello alto al livello basso: il sotto-obiettivo $\zeta^{\{l-1\}}_{\text{target}}$ definito nello spazio $Z^{\{l-1\}}$ deve essere sufficientemente vincolante da guidare la pianificazione, e la struttura di V-JEPA 2.1 garantisce che lo sia.

Si può quantificare il grado di vincolo che un sotto-obiettivo al livello $l-1$ impone sulla pianificazione al livello $l-2$. Sia $R^{\{l-1\}}(\zeta^{\{l-1\}})$ la regione degli stati primitivi $z^{\{1\}}$ compatibili con il sotto-obiettivo, definita come:

$$R^{\{l-1\}}(\zeta^{\{l-1\}}) = \{z^{\{1\}} \in Z^{\{1\}} : AGG^{\{2\}} \circ \dots \circ AGG^{\{l-1\}}(z^{\{1\}}) \approx \zeta^{\{l-1\}}\}$$

Più piccolo è il volume di $R^{\{l-1\}}(\zeta^{\{l-1\}})$ rispetto a $Z^{\{1\}}$, più il sotto-obiettivo vincola lo spazio delle soluzioni primitive. Nei sistemi JEPA con encoder equivariante, il volume di $R^{\{l-1\}}$ dipende dalla struttura delle orbite di trasformazione: due stati primitivi che differiscono per una trasformazione all'interno del gruppo G_T apprenderanno la stessa rappresentazione aggregata $\zeta^{\{l-1\}}$, quindi appartengono alla stessa classe di equivalenza sotto il sotto-obiettivo. Questo è esattamente il comportamento desiderato: il livello superiore specifica *cosa* raggiungere in termini invarianti (la categoria dell'oggetto, la sua posizione assoluta), mentre i livelli inferiori determinano *come* raggiungerlo in termini equivarianti (la traiettoria specifica, le pose intermedie).

La scalabilità temporale: dalla previsione al frame futuro alla previsione all'episodio futuro

Il beneficio computazionale della struttura gerarchica può essere quantificato analiticamente attraverso il confronto tra il costo di pianificazione piatta e il costo di pianificazione gerarchica per un dato orizzonte effettivo T_{eff} .

Per la pianificazione piatta con Cross-Entropy Method di parametri (N, M, H) : N campioni, M iterazioni, orizzonte H , il numero totale di valutazioni del World Model è:

$$C_{\text{flat}} = N \cdot M \cdot H$$

Per la pianificazione gerarchica a L livelli, ciascuno con parametri (N_l, M_l, H_l) e fattore di compressione k , l'orizzonte effettivo è $T_{\text{eff}} = \prod_{l=1}^L H_l$ (il prodotto degli orizzonti per livello), e il costo totale è:

$$C_{\text{hier}} = \sum_{l=1}^L N_l \cdot M_l \cdot H_l$$

Per $T_{\text{eff}} = H^L$ (orizzonti uguali a ogni livello), e $N_l = N$, $M_l = M$ per semplicità, si ha:

$$C_{\text{hier}} / C_{\text{flat}} = L \cdot H^1 / H^L = L / H^{L-1}$$

che è esponenzialmente più piccolo di 1 per $H > 1$ e $L > 1$. Per $H = 5$ e $L = 3$ (un sistema a tre livelli con orizzonte 5 per livello, che pianifica su un orizzonte effettivo di 125 step primitivi), si ottiene $C_{\text{hier}} / C_{\text{flat}} = 3/25$, ossia una riduzione del costo computazionale di circa un ordine di grandezza. Questo risparmio ha un costo: la pianificazione gerarchica non è ottima globalmente, perché i sotto-obiettivi selezionati dai livelli superiori possono essere sub-ottimali o addirittura irraggiungibili per i livelli inferiori. La tensione tra efficienza computazionale e qualità della pianificazione è la questione centrale dell'HRL, e il paradigma JEPA la affronta attraverso un meccanismo di *feasibility checking* che verifica se il sotto-obiettivo $\zeta^{(l-1)}_{\text{target}}$ è raggiungibile dal World Model di livello $l-1$ entro l'orizzonte H^{l-1} :

$$\zeta^{(l-1)}_{\text{target}} \text{ è fattibile sse } \min_{a_{t:t+H^{l-1}}} \|z^{(l-1)}_{t+H^{l-1}} - \zeta^{(l-1)}_{\text{target}}\|^2 \leq \epsilon_{\text{feas}}$$

dove ϵ_{feas} è una soglia di fattibilità calibrata sull'incertezza del World Model al livello $l-1$. Se il sotto-obiettivo non è fattibile, il livello l ri-campiona un'opzione alternativa — un meccanismo di feedback che previene la propagazione di sub-obiettivi irraggiungibili attraverso i livelli della gerarchia.

Proprietà di conservazione della struttura causale attraverso i livelli

Una proprietà che il framework gerarchico deve soddisfare per essere consistente con il programma del Causal Representation Learning è che la struttura causale identificata al livello l si *conservi* attraverso l'aggregazione temporale al livello $l+1$. In termini formali: se i fattori latenti $z^{(l)}$ soddisfano le condizioni di identificabilità iVAE rispetto alle azioni di livello l (condizioni C1–C4 enumerate nel Capitolo 17), è necessario che anche i fattori latenti $\zeta^{(l+1)} = \text{AGG}^{(l+1)}(z^{(l)}_{nk+1}, \dots, z^{(l)}_{(n+1)k})$ soddisfino condizioni analoghe rispetto alle opzioni di livello $l+1$.

La dimostrazione di questa conservazione è non triviale ma segue dalla struttura della composizione delle funzioni di aggregazione. Si supponga che $\text{AGG}^{(l+1)}$ sia implementato come un transformer con token $[\text{AGG}]$ che opera sulla sequenza $(z^{(l)}_{nk+1}, \omega^{(l)}_{nk}, z^{(l)}_{nk+2}, \omega^{(l)}_{nk+1}, \dots, z^{(l)}_{(n+1)k})$ di stati e azioni del livello l . Poiché i fattori

$\hat{z}^{\{l\}}$ soddisfano la condizione di distribuzione marginale gaussiana isotropa $p_{\text{SIGReg}}^{\{l\}} \approx \mathcal{N}(0, I_{\{d_l\}})$, la distribuzione dei token di input all'aggregatore è approssimativamente isotropa. Il token [AGG] prodotto dal transformer è quindi una statistica sufficiente delle azioni di livello l per predire la transizione di livello $l+1$ — la condizione di sufficienza richiesta dal teorema iVAE per l'identificabilità al livello superiore. La conservazione della struttura causale attraverso i livelli è garantita, in modo approssimativo, dalla combinazione di SIGReg a ogni livello con l'architettura transformer dell'aggregatore — una proprietà che non è codificata esplicitamente nell'obiettivo di training ma emerge dalla struttura dell'aggregazione multi-testa con attenzione sulle azioni.

La verifica empirica di questa conservazione attraverso test di probing causale applicati ai diversi livelli della gerarchia — analogamente a quanto fatto nel Capitolo 12 per i sistemi JEPA mono-livello — è uno degli obiettivi sperimentali prioritari del sotto-programma VIII del Capitolo 20, che descrive l'implementazione di un sistema gerarchico a due livelli su LeWorldModel con V-JEPA 2.1 come Perception Module.

Capitolo 10 — Model Predictive Control con JEPA come Perception Module

10.1 Il framework MPC: ottimizzazione online sulla rappresentazione latente

Il *Model Predictive Control* (MPC) è il paradigma di planning che combina un World Model con l'ottimizzazione online a orizzonte finito: ad ogni step temporale t , il sistema risolve un problema di ottimizzazione sull'orizzonte $\{t, t+1, \dots, t+H\}$, esegue soltanto la prima azione della sequenza ottimale, poi ri-ottimizza all'istante successivo con l'osservazione aggiornata. Questo ciclo — «pianifica, esegui un passo, ri-pianifica» — permette di correggere continuamente gli errori del modello e di adattarsi alle variazioni dell'ambiente.

Il problema di ottimizzazione MPC standard ha la forma:

$$a^*_{\{t:t+H\}} = \underset{a_{\{t:t+H\}}}{\operatorname{argmin}} \sum_{k=0}^{H-1} c(s_{t+k}, a_{t+k}) + c_H(\hat{s}_{t+H})$$

soggetto a: $\hat{s}_{t+k+1} = f_\theta(\hat{s}_{t+k}, a_{t+k})$ per $k = 0, \dots, H-1$

dove f_θ è il World Model, $c : (s, a) \rightarrow \mathbb{R}$ è la funzione di costo step-by-step, $c_H : s \rightarrow \mathbb{R}$ è il costo terminale, e $\hat{s}_{t+k} = \hat{s}_{t+k}(a_{t:t+k-1})$ è la traiettoria di stato prevista dal World Model date le azioni pianificate.

Il costo tipico per task robotici di manipolazione ha la struttura:

$$c(s, a) = \|x_{ee}(s) - x_{goal}\|^2 + \lambda_a \cdot \|a\|^2 + \lambda_c \cdot \mathbb{1}[\text{collision}(s)]$$

dove $x_{ee}(s)$ è la posizione dell'end-effector nello stato s , x_{goal} è la posizione target, $\|a\|^2$ è il termine di regolarizzazione sulle azioni (che penalizza movimenti bruschi), e $\mathbb{1}[\text{collision}(s)]$ è il termine di penalizzazione per collisioni.

Quando il World Model opera nello spazio latente — come nel paradigma JEPA — il problema di ottimizzazione MPC diventa:

$$\hat{z}_{t:t+H}, a_{t:t+H} = \underset{a_{t:t+H}}{\operatorname{argmin}} \sum_{k=0}^{H-1} c_z(\hat{z}_{t+k}, a_{t+k}) + c_z(z_H, \hat{z}_{t+H})$$

soggetto a: $\hat{z}_{t+k+1} = f_\theta(\hat{z}_{t+k}, a_{t+k})$ per $k = 0, \dots, H-1$ $\hat{z}_t = E_{JEPA}(o_t)$

dove E_{JEPA} è l'encoder JEPA, f_θ è il World Model nello spazio latente (che predice la prossima rappresentazione latente data la rappresentazione corrente e l'azione), e c_z è una funzione di costo definita nello spazio latente.

Il vantaggio di operare nello spazio latente di JEPA rispetto allo spazio degli stati raw è duplice. Il primo vantaggio è dimensionale: lo spazio latente JEPA ha dimensionalità molto inferiore rispetto allo spazio delle osservazioni (immagini a piena risoluzione, segnali propriocettivi, ecc.), rendendo l'ottimizzazione computazionalmente più efficiente. Il secondo vantaggio è semantico: il World Model nello spazio latente JEPA opera su una rappresentazione già pre-filtrata delle variabili irrilevanti — il pattern di rumore nei pixel, la variazione di illuminazione — e deve modellare soltanto le strutture causalmente rilevanti per la predizione.

10.2 Il World Model come predittore latente: struttura e addestramento

Il World Model nello spazio latente di JEPA ha la struttura di un predittore sequenziale: dato lo stato latente corrente $\hat{z}_t$ e l'azione a_t , predice il successivo stato latente $\hat{z}_{t+1}$:

$$\hat{z}_{t+1} = f_\theta(\hat{z}_t, a_t) + \varepsilon_t$$

dove $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_t \cdot I)$ è il rumore di predizione, e σ^2_t è la varianza del residuo — un segnale calibrato dell'incertezza del modello allo step t . In un framework deterministico, $\sigma^2_t = 0$ e il World Model è una funzione deterministica; in un framework variazionale (come in DreamerV3 o nei RSSM di Hafner), σ^2_t è parametrizzato da una rete neurale che produce l'incertezza in modo dipendente dallo stato.

L'addestramento del World Model nello spazio latente JEPA richiede di risolvere un problema di supervisione: date le rappresentazioni latenti prodotte

dall'encoder JEPA per una sequenza di frame ($\hat{z}_{t}$, $\hat{z}_{t+1}$) e l'azione a_t eseguita tra i due frame, si addestra f_θ a minimizzare la distanza tra la rappresentazione predetta e quella prodotta dall'encoder:

$$\mathcal{L}_{WM} = \mathbb{E}_{\{(o_t, a_t, o_{t+1}) \sim D\}} [\|f_\theta(E_{JEPA}(o_t), a_t) - sg(E_{JEPA}(o_{t+1}))\|^2]$$

dove $sg(\cdot)$ è lo stop-gradient — lo stesso meccanismo che impedisce il collasso rappresentazionale in I-JEPA. La presenza dello stop-gradient sul target è critica: senza di esso, il World Model e l'encoder potrebbero colludere producendo rappresentazioni triviali (tutti zero) che minimizzano la perdita senza imparare nulla di utile.

In un framework variazionale — il framework CV-JEPA proposto nel Capitolo 3 — il World Model diventa una distribuzione posteriore:

$$p_\theta(\hat{z}_{t+1} \mid \hat{z}_t, a_t) = \mathcal{N}(\mu_\theta(\hat{z}_t, a_t), \sigma^2_\theta(\hat{z}_t, a_t) \cdot I)$$

dove sia media che varianza sono parametrizzate dalla stessa rete neurale. La varianza $\sigma^2_\theta(\hat{z}_t, a_t)$ fornisce una stima calibrata dell'incertezza del World Model sull'esito dell'azione a_t a partire dallo stato $\hat{z}_t$. Questa stima di incertezza è preziosa per il MPC: permette di pesare le traiettorie pianificate in modo da preferire quelle che attraversano regioni dello spazio latente dove il World Model è più certo, evitando le regioni ad alta incertezza dove la pianificazione è inaffidabile.

10.3 Il ciclo predizione-ottimizzazione-azione nel MPC con World Model JEPA

Il ciclo MPC completo con World Model JEPA ha la seguente struttura iterativa. Al tempo t :

Passo 1 — Percezione: l'encoder JEPA produce la rappresentazione latente corrente $\hat{z}_t = E_{JEPA}(o_t)$ a partire dall'osservazione o_t .

Passo 2 — Predizione in avanti: il World Model f_θ predice la sequenza di stati latenti futuri per ciascuna traiettoria di azioni candidate. Se si usa un metodo di campionamento stocastico (CEM — Cross-Entropy Method, o MPPI — Model Predictive Path Integral), si campionano N sequenze di azioni $a^{(i)}_{t:t+H} \sim q(a_{t:t+H})$, e si calcolano le sequenze predette $\hat{z}^{(i)}_{t+1:t+H+1}$ applicando iterativamente f_θ :

$$\hat{z}^{(i)}_{t+k+1} = f_\theta(\hat{z}^{(i)}_{t+k}, a^{(i)}_{t+k}) \text{ per } k = 0, \dots, H-1$$

Passo 3 — Ottimizzazione: si valuta il costo cumulativo di ciascuna traiettoria campionata:

$$J(i) = \sum_{k=0}^{H-1} c_z(\hat{z}^{(i)}_{t+k}, a^{(i)}_{t+k}) + c_{z,H}(\hat{z}^{(i)}_{t+H})$$

e si seleziona la traiettoria ottimale $a_{t:t+H} = a^*(i^*)_{t:t+H}$ dove $i^* = \operatorname{argmin}_i J(i)$.

Passo 4 — Esecuzione e ri-pianificazione: si esegue soltanto la prima azione a^*_t della sequenza ottimale sull'ambiente reale, si osserva la nuova osservazione o_{t+1} , e si torna al Passo 1.

Il costo computazionale dominante di questo ciclo è il Passo 2: per N sequenze candidate di orizzonte H , si richiedono $N \cdot H$ valutazioni del World Model f_θ . Con un World Model leggero nello spazio latente JEPA (dimensionalità latente tipicamente 256-1024), questo calcolo è molto più efficiente che fare predizioni nello spazio delle osservazioni raw (dimensionalità milioni di pixel per frame video).

10.3.1 Il Cross-Entropy Method e il Model Predictive Path Integral come ottimizzatori MPC

L'ottimizzazione del problema MPC al Passo 3 del ciclo pianificazione-esecuzione può essere affrontata con diversi algoritmi. I due più diffusi nei World Models latenti sono il Cross-Entropy Method (CEM) e il Model Predictive Path Integral (MPPI). La loro derivazione da principi comuni chiarisce le rispettive proprietà.

Il CEM inizializza una distribuzione sulle sequenze di azioni come prodotto di distribuzioni gaussiane: $q^0(a_{t:t+H}) = \prod_{k=0}^{H-1} \mathcal{N}(a_{t+k}; \mu_k^0, \sigma_k^0 \cdot I)$. Ad ogni iterazione m , si campionano N sequenze da q^m , si valutano i costi $J(i)$, si selezionano le top- K sequenze con costo minore (l^* élite*), e si aggiorna la distribuzione come:

$$\begin{aligned} \mu_k^{m+1} &= (1/K) \sum_{i \in \text{elite}} a_{t+k}^{(i)} \\ (\sigma_k^{m+1})^2 &= (1/K) \sum_{i \in \text{elite}} (a_{t+k}^{(i)} - \mu_k^{m+1})^2 \end{aligned}$$

Dopo M iterazioni, si esegue $\mu_{\{t\}}^M$ come azione ottimale. Il CEM converge in pratica in poche iterazioni (tipicamente $M = 5-10$) per problemi di controllo di dimensionalità moderata, e non richiede il calcolo dei gradienti del World Model rispetto alle azioni — rendendolo applicabile anche quando il World Model non è differenziabile o il calcolo del gradiente è instabile.

Il MPPI, introdotto da Williams et al. (2017), ottimizza la stessa distribuzione ma attraverso una formula di aggiornamento pesata che sfrutta tutti i campioni (non solo l'élite). Si campionano N perturbazioni $\epsilon^{(i)} \sim \mathcal{N}(0, \Sigma)$ attorno alla sequenza di azioni nominale $\bar{a}_{t:t+H}$, si valutano i costi $J^{(i)} = J(\bar{a} + \epsilon^{(i)})$, e si aggiorna:

$$\bar{a}_k \leftarrow \bar{a}_k + \sum_{i=1}^N w^{(i)} \cdot \epsilon_k^{(i)}$$

dove i pesi $w^{(i)} = \exp(-J^{(i)} / \lambda) / \sum \exp(-J^{(j)} / \lambda)$ sono proporzionali alla $\exp$ del costo negativo con temperatura λ . Questo schema è equivalente al gradient ascent sulla distribuzione di importanza campionaria, ma è più stabile del gradient descent diretto per problemi con funzioni di costo discontinue o non smooth.

Per il MPC con World Model JEPA, CEM e MPPI presentano trade-off complementari. CEM è più semplice da implementare e più robusto a variazioni nell'architettura del World Model, ma richiede molti campioni ($N = 500\text{--}2000$ tipicamente) per esplorare efficacemente lo spazio delle azioni ad alta dimensionalità. MPPI richiede meno campioni ($N = 100\text{--}500$) grazie all'aggiornamento pesato che utilizza tutta l'informazione disponibile, ma richiede la scelta di una temperatura λ appropriata che dipende dalla scala del costo — un iperparametro sensibile che può richiedere calibrazione per ogni nuovo task.

10.4 - Incertezza epistemica e aleatoria nella pianificazione: derivazione del bound di subottimalità del pessimismo epistemo

Una distinzione fondamentale per il MPC con World Model è quella tra incertezza *epistemica* e incertezza *aleatoria*. L'incertezza epistemica riflette la mancanza di conoscenza del modello — è l'incertezza su f_θ stessa, non sull'ambiente. L'incertezza aleatoria riflette la stocasticità intrinseca dell'ambiente — anche con un modello perfetto, l'esito di alcune azioni è intrinsecamente variabile.

In Var-JEPA e nel framework CV-JEPA proposto nel Capitolo 3, questa distinzione emerge naturalmente dalla struttura dell'ELBO: la varianza del posteriore $q_\varphi(z|x)$ rispecchia principalmente l'incertezza aleatoria (la variabilità nell'encoding dell'osservazione), mentre la varianza della predizione $s_\psi(z_c)$ rispetto alla distribuzione $q_\varphi(z_t|x_t)$ rispecchia l'incertezza epistemica del predittore. La monografia gemella Var-T-JEPA documenta che questo segnale di incertezza è calibrato — la correlazione di Spearman con l'incertezza simulata è 0.949 su MNIST corrotto — il che suggerisce che un World Model basato su Var-JEPA potrebbe produrre stime di incertezza affidabili per guidare la pianificazione.

L'affermazione che il pianificatore debba preferire azioni verso regioni familiari in presenza di alta incertezza epistemica non è tuttavia soltanto intuitiva: è derivabile come strategia formalmente ottimale all'interno di una classe di politiche conservative, e la sua superiorità rispetto a strategie non-pessimistiche è quantificabile attraverso un bound di subottimalità che connette il tasso di errore del World Model con la perdita attesa di prestazione nella pianificazione.

Il lemma di differenza di prestazione come strumento di derivazione. Si denoti con M_{real} l'MDP reale con transizioni $p^*(z_{t+1} \mid z_t, a_t)$ e reward $r(z_t, a_t)$, e con M_θ il World Model appreso con transizioni $p_\theta(z_{t+1} \mid z_t, a_t)$ ottenute dall'encoder-predittore di Var-JEPA. Si denoti con $\eta(\pi, M)$ il rendimento cumulativo atteso di una politica π nell'ambiente M :

$$\eta(\pi, M) = \mathbb{E}_{\tau \sim \pi, M} \left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \right]$$

Il lemma di differenza di prestazione (Kakade e Langford, 2002) afferma che, per qualsiasi coppia di politiche π e qualsiasi coppia di modelli M, M' :

$$\eta(\pi, M) - \eta(\pi, M') = \frac{\gamma}{(1-\gamma)} \cdot \mathbb{E}_{\{z,a\} \sim d^\pi_{M'}} \left[\mathbb{E}_{z' \sim p_M} [V^{\pi_M}(z')] - \mathbb{E}_{z' \sim p_{M'}} [V^{\pi_{M'}}(z')] \right]$$

dove d^π_M è la misura di occupazione degli stati-azione di π in M (pesata per il fattore di sconto), e $V^{\pi_{M'}}$ è la funzione valore di π nell'ambiente M' . Il lemma stabilisce che la differenza di rendimento tra i due modelli è proporzionale alla differenza attesa nella funzione valore sotto le due distribuzioni di transizione, mediata sulla distribuzione di visite degli stati-azione di π in M_{real} .

Poiché la funzione valore è bounded da $r_{\text{max}}/(1-\gamma)$ (con r_{max} il massimo del reward per singolo step), e poiché la distanza in variazione totale tra due distribuzioni misura la differenza massima delle loro aspettative rispetto a funzioni bounded:

$$\left| \mathbb{E}_{z' \sim p^*} [V^{\pi_{M_\theta}}(z')] - \mathbb{E}_{z' \sim p_\theta} [V^{\pi_{M_\theta}}(z')] \right| \leq r_{\text{max}}/(1-\gamma) \cdot \|p^*(\cdot \mid z, a) - p_\theta(\cdot \mid z, a)\|_1$$

Sostituendo nel lemma:

$$|\eta(\pi, M_{\text{real}}) - \eta(\pi, M_\theta)| \leq \gamma \cdot r_{\text{max}} / (1-\gamma)^2 \cdot \mathbb{E}_{\{d^\pi_{M_{\text{real}}}\}} [\varepsilon(z_t, a_t)]$$

dove si è definito l'errore del modello come $\varepsilon(z_t, a_t) \triangleq \|p^*(\cdot \mid z_t, a_t) - p_\theta(\cdot \mid z_t, a_t)\|_1$. Questa disuguaglianza, che chiameremo *bound di errore del World Model*, stabilisce che la discrepanza di prestazione tra la politica ottimale nel modello appreso e la sua valutazione nell'ambiente reale è controllata dall'errore atteso del modello lungo la traiettoria seguita dalla politica stessa.

Il collegamento tra errore del World Model e incertezza epistematica di Var-JEPA. Il bound qui sopra coinvolge la distanza in variazione totale $\varepsilon(z_t, a_t)$, che non è direttamente osservabile. Var-JEPA, tuttavia, fornisce una proxy computabile per questa quantità attraverso la varianza epistematica del predittore. Nella formulazione variazionale, la distribuzione predittiva è $p_\theta(z_{t+1} \mid z_t, a_t) = \mathcal{N}(\hat{z}_{t+1}, \sigma^2_\theta(z_t, a_t) \cdot I)$ dove $\hat{z}_{t+1} = s_\psi(z_t, a_t)$ è il vettore predetto e $\sigma^2_\theta(z_t, a_t)$ è la varianza epistematica — la varianza della predizione rispetto alla distribuzione posteriore $q_\psi(z_{t+1} \mid x_{t+1})$.

In presenza di calibrazione (condizione documentata con Spearman $\rho = 0.949$ nella monografia Var-T-JEPA), esiste una costante $\kappa > 0$ tale che:

$$\varepsilon(z_t, a_t) \leq \kappa \cdot \sigma_{ep}(z_t, a_t)$$

dove $\sigma_{ep} = \sqrt{\sigma^2_{\theta}}$ è la deviazione standard epistemica. Sostituendo nel bound:

$$|\eta(\pi, M_{real}) - \eta(\pi, M_{\theta})| \leq \gamma \cdot \kappa \cdot r_{max} / (1-\gamma)^2 \cdot \mathbb{E}_{\tau \sim \pi} \{d^{\pi}\} [\sigma_{ep}(z_t, a_t)]$$

La politica pessimistica e il suo bound di subottimalità. Definito il penalized World Model $\tilde{M}$ con reward modificato $\tilde{r}(z_t, a_t) = r(z_t, a_t) - \lambda_{ep} \cdot \sigma_{ep}(z_t, a_t)$ per una costante di pessimismo $\lambda_{ep} > 0$, la politica pessimistica è:

$$\pi_{pess} = \operatorname{argmax}_{\pi} \eta(\pi, \tilde{M}) = \operatorname{argmax}_{\pi} \mathbb{E}_{\tau \sim \pi} \{ \sum_t \gamma^t (r_t - \lambda_{ep} \cdot \sigma_{ep}(z_t, a_t)) \}$$

Si può ora derivare il bound di subottimalità di π_{pess} rispetto alla politica ottimale π^* nel mondo reale. Per qualsiasi politica π :

$$\eta(\pi, M_{real}) \geq \eta(\pi, M_{\theta}) - |\eta(\pi, M_{real}) - \eta(\pi, M_{\theta})| \geq \eta(\pi, M_{\theta}) - \gamma \kappa r_{max} / (1-\gamma)^2 \cdot \mathbb{E}_{\tau \sim \pi} \{d^{\pi}\} [\sigma_{ep}]$$

Scegliendo $\lambda_{ep} = \gamma \kappa r_{max} / (1-\gamma)^2$, si ottiene:

$$\eta(\pi, M_{real}) \geq \eta(\pi, M_{\theta}) - \lambda_{ep} \cdot \mathbb{E}_{\tau \sim \pi} \{d^{\pi}\} [\sigma_{ep}] = \tilde{\eta}(\pi, \tilde{M})$$

Vale a dire: il rendimento penalizzato $\tilde{\eta}(\pi, \tilde{M})$ è un lower bound del rendimento reale $\eta(\pi, M_{real})$ per qualsiasi politica π e per la scelta canonica di λ_{ep} . Di conseguenza, ottimizzando $\tilde{\eta}$ nel modello appreso si garantisce di massimizzare un lower bound del rendimento reale.

La subottimalità di π_{pess} rispetto a π^* si deriva ora combinando questo risultato con la catena di disuguaglianze:

$$\begin{aligned} \eta(\pi^*, M_{real}) - \eta(\pi_{pess}, M_{real}) &\leq \eta(\pi^*, M_{real}) - \tilde{\eta}(\pi_{pess}, \tilde{M}) \text{ [poiché } \tilde{\eta}(\pi_{pess}) \geq \eta(\pi_{pess}, M_{real})] \\ &\leq \eta(\pi^*, M_{real}) - \tilde{\eta}(\pi^*, \tilde{M}) \text{ [ottimalità di } \pi_{pess} \text{ su } \tilde{M}] \\ &= \eta(\pi^*, M_{real}) - \eta(\pi^*, M_{\theta}) + \lambda_{ep} \cdot \mathbb{E}_{\tau \sim \pi^*} \{d^{\pi^*}\} [\sigma_{ep}] \leq 2\lambda_{ep} \cdot \mathbb{E}_{\tau \sim \pi^*} \{d^{\pi^*}\} [\sigma_{ep}] \text{ [applicando il bound di errore a } \eta(\pi^*, M_{real}) - \eta(\pi^*, M_{\theta})] \end{aligned}$$

Il bound finale è dunque:

$$\eta(\pi^*, M_{real}) - \eta(\pi_{pess}, M_{real}) \leq 2\gamma \kappa r_{max} / (1-\gamma)^2 \cdot \mathbb{E}_{\tau \sim \pi^*} \{d^{\pi^*}\} [\sigma_{ep}(z_t, a_t)]$$

Il lato destro è l'incertezza epistemica attesa lungo la traiettoria della politica ottimale π^* , non lungo la traiettoria di π_{pess} . Questo è un risultato di rilievo

concettuale: la subottimalità di π_{pess} dipende non da quanto π_{pess} sia incerta nelle regioni che visita (che sono regioni familiari per costruzione), ma da quanto π^* — la politica che si vorrebbe ma non si può eseguire senza rischio di errore del modello — visitasse regioni ad alta incertezza epistemica.

Conseguenze per il MPC con World Model basato su Var-JEPA. Il bound derivato formalizza tre implicazioni operative. Primo: la strategia pessimistica è non soltanto intuitivamente motivata ma formalmente ottimale (nell'accezione di massimizzare un lower bound garantito), con una costante $\lambda_{\text{ep}} = \gamma r_{\text{max}} / (1-\gamma)^2$ che è direttamente calibrabile dai parametri del sistema. Secondo: la qualità del bound dipende criticamente dalla calibrazione κ del segnale di incertezza di Var-JEPA — un modello mal calibrato (κ grande) rende λ_{ep} eccessivamente conservativo, penalizzando azioni che in realtà sarebbero sicure. Il risultato $\rho = 0.949$ di Var-T-JEPA suggerisce che la calibrazione è sufficientemente accurata per rendere il bound stretto. Terzo: il bound si deteriora proporzionalmente al fattore $\gamma \cdot r_{\text{max}} / (1-\gamma)^2$ — per orizzonti lunghi ($\gamma \rightarrow 1$) o reward di ampia variabilità (r_{max} grande), la penalizzazione epistemica deve essere proporzionalmente maggiore, il che si traduce in un pianificatore più conservativo.

Questa derivazione connette la distinzione epistemica/aleatoria introdotta all'inizio della sezione con il corpus di risultati dell'offline reinforcement learning (Levine et al. 2020; Kumar et al. 2020; Kidambi et al. 2020): il pessimismo epistemico nel MPC con World Model è formalmente analogo alla penalizzazione conservativa delle azioni fuori distribuzione nel CQL di Kumar et al. e nel MOPO di Kidambi et al., con la differenza che la nozione di «fuori distribuzione» è qui direttamente quantificata attraverso la varianza epistemica del predittore latente di Var-JEPA, anziché attraverso una stima della densità nello spazio osservazione-azione.

Capitolo 11 — Reinforcement Learning gerarchico con World Models basati su JEPA

11.1 La struttura del RL gerarchico: policy a livelli multipli

Il *Reinforcement Learning gerarchico* (HRL) decompone il problema di apprendimento del controllo in una gerarchia di policy che operano su timescale diverse. Nel framework di Sutton et al. (1999) delle *opzioni* — la formalizzazione più influente dell'HRL — si definisce un'opzione ω come una tripla $(I_\omega, \pi_\omega, \beta_\omega)$ dove $I_\omega \subseteq S$ è l'insieme degli stati in cui l'opzione può essere avviata, $\pi_\omega : S \times A \rightarrow [0,1]$ è la policy intra-opzione che determina le azioni primitive durante l'esecuzione, e $\beta_\omega : S \rightarrow [0,1]$ è la funzione di terminazione che determina quando l'opzione termina.

Un sistema HRL a due livelli opera come segue: la *meta-policy* $\pi_{\text{high}} : S \rightarrow \Omega$ seleziona un'opzione $\omega_t \in \Omega$ ad ogni intervallo di k step primitivi; l'opzione selezionata esegue la sua policy intra-opzione π_{ω} per un numero di step determinato dalla funzione di terminazione β_{ω} . La meta-policy ottimizza il return cumulativo sull'orizzonte lungo; ciascuna opzione ottimizza un sotto-obiettivo locale che contribuisce al reward globale.

Nel contesto dei World Models basati su JEPA, questa struttura si arricchisce di un modello delle dinamiche a ciascun livello gerarchico. Al livello basso, il World Model latente $f_{\text{low}} : (\hat{z}_t, a_t) \rightarrow \hat{z}_{t+1}$ predice la transizione tra stati latenti alta frequenza. Al livello alto, il World Model astratto $f_{\text{high}} : (\zeta_k, \omega_k) \rightarrow \zeta_{k+1}$ predice la transizione tra stati latenti a bassa frequenza (ogni k step primitivi), condizionato sull'opzione selezionata ω_k . Gli stati latenti ad alto livello ζ sono ottenuti aggregando temporalmente le rappresentazioni JEPA: $\zeta_k = \text{agg}(\hat{z}_{tk}, \hat{z}_{tk+1}, \dots, \hat{z}_{tk+k-1})$ dove $\text{agg}(\cdot)$ è una funzione di aggregazione (media, max-pooling, o un'altra RNN leggera).

11.2 La funzione di reward intrinseco come segnale di apprendimento gerarchico

Una sfida fondamentale dell'HRL è come definire il reward per le opzioni di basso livello: il reward del task globale è tipicamente sparso — arriva soltanto al completamento del task — e non fornisce un segnale di addestramento sufficiente per le opzioni di basso livello.

Il framework JEPA suggerisce una risposta naturale: usare la *predittività della rappresentazione latente* come reward intrinseco per le opzioni di basso livello. L'idea è che un'opzione che porta il sistema verso stati latenti più facilmente predicibili dal World Model fornisce un contributo positivo all'apprendimento globale: esegue le azioni che costruiscono rappresentazioni strutturate del mondo, e queste rappresentazioni strutturate sono la condizione necessaria per la pianificazione efficace ad alto livello.

Formalmente, il reward intrinseco al livello basso è:

$$r_{\text{intr}}(\hat{z}_t, a_t, \hat{z}_{t+1}) = - \|\hat{z}_{t+1} - f_{\theta}(\hat{z}_t, a_t)\|^2 / \sigma_t^2$$

dove il denominatore σ_t^2 normalizza il segnale in funzione dell'incertezza del World Model: le azioni in regioni dello spazio latente dove il World Model è già incerto producono un reward intrinseco basso (il World Model sa che non sa), mentre le azioni che portano in regioni inaspettate ma predicibili producono un reward intrinseco alto. Questa forma di reward è collegata al *curiosity-driven exploration* di Pathak et al. (2017) e al *disagreement-based exploration* di Pathak et al. (2019), ma si distingue da entrambi per l'uso dello spazio latente JEPA anziché dello spazio delle feature di un encoder separato.

Il reward complessivo per la policy di basso livello è una combinazione pesata di reward estrinseco (fornito dal task) e reward intrinseco:

$$r_{\text{low}}(\hat{z}_t, a_t, \hat{z}_{t+1}) = r_{\text{ext}}(\hat{z}_{t+1}) + \lambda_{\text{intr}} \cdot r_{\text{intr}}(\hat{z}_t, a_t, \hat{z}_{t+1})$$

dove λ_{intr} è un coefficiente di bilanciamento che può essere adattato dinamicamente nel corso del training: all'inizio dell'apprendimento, quando il reward estrinseco è raramente raggiunto, λ_{intr} è alto per guidare l'esplorazione; man mano che il sistema impara a raggiungere il goal, λ_{intr} può essere ridotto per concentrare l'ottimizzazione sul reward estrinseco.

11.2.1 Connessione tra reward intrinseco JEPa e information gain

Il reward intrinseco proposto $r_{\text{intr}}(\hat{z}_t, a_t, \hat{z}_{t+1}) = -\|\hat{z}_{t+1} - f_{\theta}(\hat{z}_t, a_t)\|^2 / \sigma^2_{\theta}$ ha una connessione formale con la nozione di *information gain* bayesiana che vale la pena esplicitare, poiché illumina i casi in cui il reward è massimo e quelli in cui è zero.

In un World Model variazionale (come quello derivato nel Capitolo 10), la distribuzione predittiva è $p_{\theta}(\hat{z}_{t+1} \mid \hat{z}_t, a_t) = \mathcal{M}(f_{\theta}(\hat{z}_t, a_t), \sigma^2_{\theta}(\hat{z}_t, a_t) \cdot \mathbf{I})$. Il log-likelihood del successore osservato $\hat{z}_{t+1}$ sotto questo modello è:

$$\log p_{\theta}(\hat{z}_{t+1} \mid \hat{z}_t, a_t) = -\|\hat{z}_{t+1} - f_{\theta}(\hat{z}_t, a_t)\|^2 / (2\sigma^2_{\theta}) - (d/2)\log(2\pi\sigma^2_{\theta})$$

Il reward intrinseco r_{intr} è quindi proporzionale al negativo del log-likelihood — ovvero al log-likelihood *negativo* del successore osservato sotto il modello. Azioni che portano in regioni dello spazio latente dove il World Model fa previsioni accurate (σ^2_{θ} piccola e f_{θ} vicino a $\hat{z}_{t+1}$) producono log-likelihood alto e quindi reward intrinseco basso. Azioni che portano in regioni dove il World Model è sorpreso — dove la previsione è lontana dall'osservazione reale — producono log-likelihood basso e reward intrinseco alto. Il sistema è quindi spinto a esplorare regioni dove il World Model non sa ancora fare previsioni accurate.

Questa interpretazione lega il reward intrinseco all'*information gain* epistemico: le azioni che producono transizioni sorprendenti per il World Model sono quelle che forniscono la maggiore informazione per aggiornare i parametri θ , poiché l'errore di previsione è esattamente il segnale usato per aggiornare il World Model durante il training. Il coefficiente λ_{intr} che scala il reward intrinseco ha quindi un'interpretazione bayesiana: è proporzionale al tasso a cui il sistema valuta il guadagno informativo rispetto al reward estrinseco. All'inizio del training, quando il World Model è ancora molto incerto ovunque, λ_{intr} alto incentiva l'esplorazione di nuove regioni per raccogliere dati informativi; a convergenza, quando il World Model è accurato nelle regioni rilevanti per il task, λ_{intr} basso concentra l'ottimizzazione sul reward estrinseco.

Derivazione del legame tra il coefficiente λ_{intr} e la velocità di convergenza del World Model

L'interpretazione bayesiana di λ_{intr} proposta nel paragrafo precedente — quel coefficiente è proporzionale al tasso a cui il sistema valuta il guadagno informativo rispetto alla ricompensa estrinseca — ha la struttura di un'analogia qualitativa più che di un risultato derivabile. Essa stabilisce un'intuizione corretta ma lascia aperta la domanda più operativamente rilevante: come si traduce una scelta specifica di λ_{intr} in una traiettoria specifica di riduzione dell'errore del World Model, e quale valore di λ_{intr} garantisce la convergenza più rapida dati gli obiettivi del task? La risposta a questa domanda richiede di formalizzare la velocità di convergenza del World Model come funzione del profilo di esplorazione indotto dal coefficiente, e di derivare il valore ottimale adattivo $\lambda^*_{\text{intr}}(t)$ che minimizza l'errore di predizione all'istante t mantenendo compatibilità con il task estrinseco. Si tratta di un collegamento che rende operativa la calibrazione adattiva proposta nel testo, trasformandola da principio euristico a regola di aggiornamento derivabile.

La velocità di convergenza del World Model come funzione della copertura. Sia $f^*(z, a) = \mathbb{E}[z' \mid z, a]$ la funzione di transizione vera nello spazio latente $\mathbb{Z} = \mathbb{R}^{\mathcal{D}_Z}$, e sia $f_\theta: \mathbb{Z} \times \mathcal{A} \rightarrow \mathbb{Z}$ il World Model parametrico addestrato per minimizzare l'errore quadratico medio sulla distribuzione di transizioni raccolta dall'agente. Si definisca l'errore globale di predizione del World Model al tempo di addestramento t come:

$$\epsilon_{\text{WM}}(t) = \mathbb{E}_{(z, a) \sim \mathcal{Q}_{\text{task}}} [\|f_\theta(z, a) - f^*(z, a)\|^2]$$

dove $\mathcal{Q}_{\text{task}}$ è la distribuzione marginale sulle coppie (z, a) rilevanti per il task — la distribuzione che determina le prestazioni del sistema sull'obiettivo estrinseco e che, in un task di manipolazione robotica, è concentrata sulle configurazioni raggiungibili dell'oggetto target. L'errore $\epsilon_{\text{WM}}(t)$ è la quantità che il World Model deve ridurre per garantire che la pianificazione immaginata (il rollout latente della sezione 11.5) produca traiettorie fedeli all'evoluzione reale del sistema.

Per un World Model con funzione di transizione Lipschitz — ossia tale che $\|f^*(z_1, a) - f^*(z_2, a)\| \leq L_f \cdot \|z_1 - z_2\|$ per ogni (z_1, z_2, a) — l'errore di predizione in un punto (z, a) non visitato durante il training è maggiorato da:

$$\|f_\theta(z, a) - f^*(z, a)\|^2 \leq L_f^2 \cdot \min_{(z_i, a_i) \in \mathcal{D}_t} \|(z, a) - (z_i, a_i)\|^2 + \sigma^2_{\text{noise}} / |\mathcal{D}_t|$$

dove il primo addendo è il termine di copertura — misura quanto densamente le transizioni campionate in $\mathcal{D}_t = \{(z_i, a_i, z'_i)\}_{i=1}^{N_t}$ coprono lo spazio rilevante — e il secondo è il termine di varianza statistica, che decresce con la radice del numero di campioni indipendentemente dalla loro distribuzione nello spazio.

Prendendo il valore atteso rispetto a $\mathcal{Q}_{\text{task}}$:

$$\epsilon_{\text{WM}}(t) \leq L_{\text{f}}^2 \cdot \mathbb{E}_{\{(z,a) \sim \mathcal{Q}_{\text{task}}\}} [\min_{\{i \leq N_{\text{t}}\}} \|(z, a) - (z_i, a_i)\|^2] + \sigma_{\text{noise}}^2 / N_{\text{t}}$$

Il primo addendo — che si può denotare $\Delta_{\text{cov}}(t)$ — è la *copertura residua* del task-space da parte del dataset $\mathcal{D}_{\text{t}}$. È questa quantità che dipende crucialmente dalla politica di esplorazione adottata e, quindi, dal valore di λ_{intr} .

La politica di esplorazione indotta da λ_{intr} e il suo effetto sulla copertura. Il reward composito che guida la raccolta dati è:

$$r_{\text{low}}(z, a) = r_{\text{ext}}(z, a) + \lambda_{\text{intr}} \cdot r_{\text{intr}}(z, a)$$

con $r_{\text{intr}}(z, a) = -\|\hat{z}' - f_{\theta}(\hat{z}, a)\|^2 / \sigma_{\text{t}}^2$. Poiché r_{intr} è massimo nelle regioni dove il World Model fa errori di predizione più grandi, una politica che massimizza il valore cumulativo sotto r_{low} è spinta a visitare preferenzialmente le regioni ad alta incertezza del World Model quando λ_{intr} è grande. In termini della distribuzione di occupazione $\mu_{\{\lambda\}}(z, a)$ indotta dalla politica ottimale sotto reward composito con coefficiente λ_{intr} :

$$\mu_{\{\lambda\}}(z, a) \propto \exp(\beta \cdot (V_{\text{ext}}(z) + \lambda_{\text{intr}} \cdot V_{\text{intr}}(z))) \cdot \mathcal{Q}_0(z, a)$$

dove β è il coefficiente di temperatura della policy (che regola il grado di ottimalità), V_{ext} e V_{intr} sono le funzioni di valore per i due componenti del reward, e $\mathcal{Q}_0$ è la distribuzione stazionaria del processo di Markov sotto la politica di base. L'entropia della distribuzione di occupazione è strettamente crescente in λ_{intr} nella misura in cui V_{intr} ha gradiente non nullo nelle regioni inesplorate: un λ_{intr} più elevato «appiattisce» la distribuzione di occupazione verso regioni a basso valore estrinseco ma alta incertezza predittiva, producendo una copertura più uniforme dello spazio delle transizioni.

Per un'approssimazione di primo ordine nelle regioni dove V_{intr} domina V_{ext} — ossia nelle regioni inesplorate dove il World Model ha errori grandi — la distribuzione di occupazione adicional è proporzionale a $\exp(\beta \cdot \lambda_{\text{intr}} \cdot V_{\text{intr}}(z))$, il che significa che la probabilità di visitare una regione con incertezza $\epsilon_{\text{loc}}(z) = \|\hat{z}' - f_{\theta}(\hat{z})\|^2$ è:

$$P_{\{\mu_{\lambda}\}}(z) \propto \exp(\beta \cdot \lambda_{\text{intr}} \cdot \epsilon_{\text{loc}}(z) / \sigma_{\text{t}}^2)$$

Questa è la distribuzione di Boltzmann sullo spazio delle incertezze del World Model, con temperatura $\tau_{\text{exploration}} = \sigma_{\text{t}}^2 / (\beta \cdot \lambda_{\text{intr}})$. Quando $\tau_{\text{exploration}} \rightarrow 0$ (ossia $\lambda_{\text{intr}} \rightarrow \infty$), l'agente visita deterministicamente le regioni di massima incertezza; quando $\tau_{\text{exploration}} \rightarrow \infty$ (ossia $\lambda_{\text{intr}} \rightarrow 0$), l'agente ignora completamente l'incertezza e massimizza solo il reward estrinseco.

La derivazione della velocità di convergenza. Dati N_t campioni estratti dalla distribuzione di occupazione $\mu_{\{\lambda\}}$, la copertura residua $\Delta_{\text{cov}}(t)$ può essere stimata attraverso la teoria della covering number. Per un task-space di dimensione effettiva d_{eff} con raggio R — si assume che q_{task} sia concentrata su una varietà di dimensione d_{eff} nello spazio di stato-azione — il numero minimo di punti necessari a coprire la varietà a risoluzione δ è dell'ordine $(R/\delta)^{d_{\text{eff}}}$. La copertura residua dopo N_t campioni estratti da $\mu_{\{\lambda\}}$ è:

$$\Delta_{\text{cov}}(t) \approx C_{\text{geo}} \cdot R^2 \cdot (d_{\text{eff}} / N_t)^{2/d_{\text{eff}}} \cdot (\sigma_t^2 / (\beta \cdot \lambda_{\text{intr}}))^{2/d_{\text{eff}}}$$

dove C_{geo} è una costante geometrica dipendente dalla curvatura della varietà su cui q_{task} è concentrata, e il fattore $(\sigma_t^2 / (\beta \cdot \lambda_{\text{intr}}))^{2/d_{\text{eff}}}$ riflette il fatto che una distribuzione di occupazione più uniforme — corrispondente a λ_{intr} più grande — riduce la copertura residua più rapidamente. Sostituendo nella maggiorazione dell'errore:

$$\varepsilon_{\text{WM}}(t) \leq L_f^2 \cdot C_{\text{geo}} \cdot R^2 \cdot (d_{\text{eff}} / N_t)^{2/d_{\text{eff}}} \cdot (\sigma_t^2 / (\beta \cdot \lambda_{\text{intr}}))^{2/d_{\text{eff}}} + \sigma_{\text{noise}}^2 / N_t$$

Poiché $N_t = t \cdot \nu$, dove ν è il tasso di campionamento dell'ambiente (transizioni per unità di tempo di addestramento), la velocità di decrescita dell'errore è:

$$d\varepsilon_{\text{WM}} / dt \approx -(2/d_{\text{eff}}) \cdot L_f^2 \cdot C_{\text{geo}} \cdot R^2 \cdot \nu^{2/d_{\text{eff}}} \cdot (d_{\text{eff}})^{2/d_{\text{eff}}} \cdot (\sigma_t^2 / (\beta \cdot \lambda_{\text{intr}}))^{2/d_{\text{eff}}} \cdot t^{-(2/d_{\text{eff}} + 1)} - \sigma_{\text{noise}}^2 \cdot \nu / (\nu \cdot t)^2$$

Il termine dominante nella velocità di convergenza è proporzionale a:

$$d\varepsilon_{\text{WM}} / dt \sim -K(\lambda_{\text{intr}}) \cdot t^{-(2 + d_{\text{eff}})/d_{\text{eff}}}$$

dove la costante $K(\lambda_{\text{intr}}) = L_f^2 \cdot C_{\text{geo}} \cdot R^2 \cdot \nu^{2/d_{\text{eff}}} \cdot (\beta \cdot \lambda_{\text{intr}} / \sigma_t^2)^{2/d_{\text{eff}}}$ è strettamente crescente in λ_{intr} . Questa relazione formale esprime il risultato centrale: un valore più elevato del coefficiente di ricompensa intrinseca accelera la convergenza del World Model secondo una legge di potenza, con esponente di velocità $(2 + d_{\text{eff}})/d_{\text{eff}}$ che cresce al decrescere della dimensionalità effettiva del task-space. Per task robotici di manipolazione con oggetti rigidi, d_{eff} è tipicamente nell'intervallo 6–12 (sei gradi di libertà dell'oggetto più eventuali variabili di configurazione), il che produce esponenti di convergenza nell'intervallo $(2 + 6)/6 = 4/3$ e $(2 + 12)/12 = 7/6$. La convergenza è sub-lineare nel tempo e accelerata da λ_{intr} in modo sub-lineare nell'esponente.

Il costo dell'esplorazione eccessiva: la degradazione della policy estrinseca. La relazione $K(\lambda_{\text{intr}})$ crescente mostrerebbe, se considerata in isolamento, che scegliere λ_{intr} il più grande possibile accelera sempre la convergenza del World Model. Questa conclusione è però incompleta perché

ignora il costo estrinseco dell'esplorazione: un agente che dedica una frazione crescente del suo comportamento all'esplorazione delle regioni ad alta incertezza si allontana dalla distribuzione ϱ_{task} rilevante per il task, raccogliendo meno campioni informativi per il task stesso e degradando la policy estrinseca.

Il valore estrinseco atteso della policy sotto reward composito con coefficiente λ_{intr} è:

$$J_{\text{ext}}(\lambda_{\text{intr}}) = \mathbb{E}_{\pi_{\lambda_{\text{intr}}}} \left[\sum_{t=0}^T \gamma^t \cdot r_{\text{ext}}(z_t, a_t) \right]$$

Questa quantità è strettamente decrescente in λ_{intr} per λ_{intr} sufficientemente grande, perché la distribuzione di occupazione μ_{λ} si allontana da ϱ_{task} e la policy raccoglie ricompense estrinseche sub-ottimali durante la fase di esplorazione. La degradazione può essere quantificata attraverso la perdita di performance della policy:

$$J_{\text{ext}}(0) - J_{\text{ext}}(\lambda_{\text{intr}}) \leq \lambda_{\text{intr}} \cdot \|V_{\text{intr}}\|_{\infty} \cdot D\{\text{TV}\}(\mu_0, \mu_{\lambda}) / (1 - \gamma)$$

dove $D\{\text{TV}\}$ è la distanza di variazione totale tra le distribuzioni di occupazione con e senza reward intrinseco, e $\|V_{\text{intr}}\|_{\infty}$ è l'intensità massima del segnale esplorativo. Il bound è proporzionale a λ_{intr} , confermando che la degradazione estrinseca cresce linearmente con il coefficiente di esplorazione nel regime di λ_{intr} moderato.

Il coefficiente ottimale adattivo e la sua traiettoria temporale. Il bilancio tra la riduzione dell'errore del World Model e la degradazione della policy estrinseca è catturato dall'obiettivo composito:

$$J_{\text{total}}(\lambda_{\text{intr}}, t) = J_{\text{ext}}(\lambda_{\text{intr}}) - c \cdot \epsilon_{\text{WM}}(t; \lambda_{\text{intr}})$$

dove c è un peso che traduce l'errore del World Model in unità commensurabili con la ricompensa estrinseca. Ottimizzando rispetto a λ_{intr} al tempo t , la condizione del primo ordine — che uguaglia il guadagno marginale sulla riduzione dell'errore del World Model con la perdita marginale sulla ricompensa estrinseca — produce:

$$\partial J_{\text{ext}} / \partial \lambda_{\text{intr}} + c \cdot (2 / (d_{\text{eff}} \cdot \sigma^2_t)) \cdot K_0 \cdot (\beta \cdot \lambda_{\text{intr}} / \sigma^2_t)^{2/d_{\text{eff}} - 1} \cdot (\beta / \sigma^2_t) \cdot t^{-2/d_{\text{eff}}} = 0$$

dove $K_0 = L_f^2 \cdot C_{\text{geo}} \cdot R^2 \cdot v^{2/d_{\text{eff}}} \cdot d_{\text{eff}}^{2/d_{\text{eff}}}$. Risolvendo rispetto a λ_{intr} e semplificando i termini costanti:

$$\lambda_{\text{intr}}^*(t) = (\sigma^2_t / \beta) \cdot \left[|\partial J_{\text{ext}} / \partial \lambda_{\text{intr}}|_{\lambda=0} \cdot \sigma^2_t / (c \cdot (2/d_{\text{eff}}) \cdot K_0 \cdot \beta) \right]^{d_{\text{eff}} / (2 - d_{\text{eff}})} \cdot t^{d_{\text{eff}} / (d_{\text{eff}} - 2) \cdot 2/d_{\text{eff}}}$$

Per la semplificazione rilevante — il caso in cui $\partial J_{\text{ext}} / \partial \lambda_{\text{intr}} \approx -\alpha_{\text{cost}}$ per una costante α_{cost} che esprime la perdita estrinseca marginale per unità di esplorazione — la traiettoria ottimale si riduce a:

$$\lambda^*_{\text{intr}}(t) = A \cdot (\sigma^2_t)^{\{(d_{\text{eff}} + 2)/(d_{\text{eff}})\}} \cdot t^{\{-2/(d_{\text{eff}} - 2)\}} \cdot (\alpha_{\text{cost}} / (c \cdot K_0))^{\{d_{\text{eff}} / (2 - d_{\text{eff}})\}}$$

dove A è una costante numerica che assorbe i termini di β e gli esponenti. Il comportamento asintotico è:

$$\lambda^*_{\text{intr}}(t) \sim A' \cdot \sigma^2_t \cdot t^{\{-2/(d_{\text{eff}} - 2)\}} \text{ per } t \rightarrow \infty, \text{ con } d_{\text{eff}} > 2$$

La decrescita di $\lambda^*_{\text{intr}}(t)$ con il tempo è la formalizzazione del principio qualitativo enunciato nel testo: a convergenza, quando il World Model è accurato nelle regioni rilevanti per il task, l'esplorazione ottimale tende a zero. Ma la formula derivata aggiunge l'informazione quantitativa cruciale: la velocità di tale decrescita dipende dalla dimensionalità effettiva del task-space d_{eff} . Per task a bassa dimensionalità effettiva (d_{eff} vicino a 2), l'esponente $t^{\{-2/(d_{\text{eff}} - 2)\}}$ diverge e la calibrazione adattiva richiede una riduzione molto rapida di λ_{intr} ; per task ad alta dimensionalità ($d_{\text{eff}} \gg 2$), l'esponente si avvicina a zero e λ_{intr} deve rimanere significativo per un tempo proporzionalmente più lungo.

La connessione con la varianza σ^2_t del World Model. Un'osservazione supplementare di rilievo pratico emerge dall'analisi del fattore σ^2_t nella traiettoria $\lambda_{\text{intr}}(t)$. In un World Model variazionale — come quello derivato nel Capitolo 10, in cui la distribuzione predittiva è $p_{\theta}(\tilde{z}_{t+1} \mid \hat{z}_t, a_t) = \mathcal{M}(f_{\theta}(\hat{z}_t, a_t), \sigma_{\theta}(\hat{z}_t, a_t) \cdot I)$ —, la varianza σ^2_t non è fissa ma è essa stessa stimata dal modello e decresce durante il training nella misura in cui il modello diventa più certo delle proprie predizioni. La relazione $\lambda_{\text{intr}}(t) \sim A' \cdot \sigma^2_t \cdot t^{\{-2/(d_{\text{eff}} - 2)\}}$ suggerisce allora una regola di calibrazione adattiva automatica:

$$\lambda_{\text{intr}}(t) \leftarrow \eta \cdot \sigma^2_{\theta}(t) \cdot \mathbb{E}_{\{z \sim \mu_{\text{task}}\}}[\sigma^2_{\theta}(z, a)]$$

dove η è un iperparametro di scala e il valore atteso è stimato attraverso campionamento dalla distribuzione di occupazione corrente. Questa regola ha la proprietà di auto-regolarsi senza richiedere schedule prefissate: quando il World Model è molto incerto — $\sigma^2_{\theta}(t)$ grande —, λ_{intr} è automaticamente alto e l'agente esplora più aggressivamente; quando il World Model converge — $\sigma^2_{\theta}(t) \rightarrow 0$ — λ_{intr} si azzera e la policy si concentra esclusivamente sull'obiettivo estrinseco. Il meccanismo è esattamente il collegamento tra la velocità di convergenza del World Model e la calibrazione ottimale del coefficiente di ricompensa intrinseca che rendeva tale calibrazione «operativa» soltanto a livello di principio nel paragrafo precedente, e che ora è espressa come regola computabile a partire dalle stime di varianza prodotte direttamente dal modello variazionale.

11.3 Apprendimento separato di encoder e World Model nel paradigma JEPa

Una proprietà architettonica particolarmente vantaggiosa del paradigma JEPa — rispetto al paradigma DreamerV3 o RSSM di Hafner — è la possibilità di

separare l'addestramento dell'encoder visivo dall'addestramento del World Model. In DreamerV3, l'encoder e il World Model sono addestrati congiuntamente attraverso l'ELBO del RSSM, il che significa che le rappresentazioni latenti prodotte dall'encoder sono specificamente ottimizzate per la predittività del World Model, ma possono non essere le rappresentazioni più utili per il controllo o per il trasferimento a nuovi ambienti.

In un sistema gerarchico con Perception Module JEPA, l'encoder E_{JEPA} è preaddestrato indipendentemente su grandi corpora di osservazioni (immagini, video) con l'obiettivo JEPA, producendo rappresentazioni con le proprietà di invarianza e struttura causale discusse nelle Parti Prima e Seconda. Successivamente, il World Model f_θ e le policy gerarchiche vengono addestrate *a encoder congelato*, usando le rappresentazioni latenti come feature fisse.

Questa separazione produce tre vantaggi. Il primo è la *riusabilità*: lo stesso encoder JEPA preaddestrato può essere usato per World Models in ambienti diversi senza dover ri-addestrare la componente percettiva. Il secondo è la *stabilità*: addestrare il World Model su rappresentazioni fisse evita i problemi di instabilità che emergono quando encoder e modello si adattano reciprocamente — un problema documentato nella letteratura del model-based RL. Il terzo è la *scalabilità del preaddestramento*: l'encoder può essere preaddestrato su ordini di grandezza più dati di quelli disponibili per il task specifico — video youtube, dati egocentrici, immagini satellite — catturando strutture visive universali che supportano il learning rapido del World Model su dati task-specifici.

La limitazione di questo approccio è il rischio che le rappresentazioni JEPA preaddestrate non siano ottimali per il World Model specifico che viene addestrato su di esse: le caratteristiche che JEPA privilegia — invarianza alle variazioni superficiali, predittività nello spazio latente — possono non corrispondere perfettamente alle caratteristiche più rilevanti per la funzione di transizione del task. Questo compromesso tra universalità del Perception Module e specificità del World Model è una delle tensioni fondamentali del paradigma AMI che la ricerca futura dovrà risolvere.

11.4 Connessione con DreamerV3 e la famiglia RSSM

La famiglia di World Models sviluppata da Danijar Hafner — RSSM (2018), DreamerV1-V3 (2020-2023) — rappresenta il punto di riferimento più maturo per l'apprendimento di World Models latenti con RL. Il Recurrent State Space Model (RSSM) di Hafner modella lo stato latente come la composizione di una componente deterministica h_t (prodotta da un RNN) e una componente stocastica z_t (campionata da una distribuzione gaussiana condizionata su h_t):

$h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1})$ [componente deterministica — GRU] $z_t \sim p_\psi(z_t | h_t)$ [componente stocastica — prior] $z_t \sim q_\psi(z_t | h_t, x_t)$ [posteriore]

dove x_t è l'osservazione e l'encoder non-lineare della RSSM è il corrispettivo dell'encoder JEPA nella formulazione alternativa proposta in questa monografia.

La differenza fondamentale tra l'RSSM di DreamerV3 e un World Model basato su Var-JEPA è nell'obiettivo di addestramento dell'encoder. DreamerV3 addestra l'encoder congiuntamente con il World Model attraverso l'ELBO dell'RSSM:

$$\mathcal{L}_{\text{DreamerV3}} = \mathbb{E}_{\{q\}}[\log p_{\theta}(x_t | h_t, z_t)] - D_{\text{KL}}(q_{\psi}(z_t | h_t, x_t) \parallel p_{\psi}(z_t | h_t))$$

Il termine di ricostruzione $\log p_{\theta}(x_t | h_t, z_t)$ obbliga il modello a mantenere informazione sufficiente per ricostruire l'osservazione — il che, come argomenta LeCun, può portare a spendere capacità rappresentativa sulla modellazione di dettagli percettivi irrilevanti per il controllo.

Var-JEPA, al contrario, sostituisce il termine di ricostruzione con il termine di predittività dello spazio latente — l'ELBO a cinque termini che non include mai la ricostruzione delle osservazioni grezze. Questo significa che le rappresentazioni latenti di un World Model basato su Var-JEPA sono strutturate per la predittività intra-spazio-latente, non per la ricostruibilità delle osservazioni. L'ipotesi — non ancora verificata in modo sistematico — è che questo produca World Models con migliore efficienza di campionamento e migliore generalizzazione, al prezzo di una valutazione qualitativa più difficile (non si può «guardare» cosa il modello ha imparato a predire, perché non produce immagini ricostruite).

DreamerV3 ha dimostrato su benchmarks diversificati — Atari, DMControl, Crafter, BSuite — prestazioni state-of-the-art con un singolo set di iperparametri. Se un World Model basato su Var-JEPA possa eguagliare o superare questo risultato è una delle domande empiriche più rilevanti che il programma di ricerca descritto in questa monografia dovrà rispondere.

11.4.1 Confronto formale tra l'ELBO dell'RSSM di DreamerV3 e l'ELBO di Var-JEPA: statistiche sufficienti, geometria della rappresentazione e condizioni di equivalenza

La sezione 11.4 ha posto la questione se un World Model basato su Var-JEPA possa eguagliare le prestazioni di DreamerV3 — la domanda empirica più rilevante del capitolo. La risposta richiede un'analisi formale delle due famiglie di modelli in termini di ciò che ciascuna ottimizza, di quali sono le statistiche sufficienti prodotte dai rispettivi encoder, e in quale regime le due famiglie convergono verso lo stesso spazio di soluzioni. Senza questa analisi, il confronto empirico rischia di essere non-interpretabile: differenze di performance potrebbero riflettere differenze architetturali irrilevanti piuttosto che differenze strutturali negli obiettivi.

L'ELBO dell'RSSM. In DreamerV3, lo stato latente è la coppia (h_t, z_t) dove h_t è la componente deterministica (GRU) e z_t è la componente stocastica (categorica o gaussiana). L'ELBO dell'RSSM, che DreamerV3 massimizza rispetto ai parametri dell'encoder $q_\varphi(z_t | h_t, x_t)$, del prior $p_\psi(z_t | h_t)$ e del decoder $p_\theta(x_t | h_t, z_t)$, è:

$$\mathcal{L}_{\text{RSSM}} = \mathbb{E}_{q_\varphi} [\sum_t \log p_\theta(x_t | h_t, z_t)] - \beta \cdot \sum_t D_{\text{KL}}(q_\varphi(z_t | h_t, x_t) \parallel p_\psi(z_t | h_t))$$

Il primo termine è il log-likelihood della ricostruzione: l'encoder è costretto a mantenere nelle rappresentazioni (h_t, z_t) tutta l'informazione necessaria a ricostruire x_t . Il secondo termine è la penalità KL verso il prior: le rappresentazioni non devono essere troppo lontane dalla distribuzione predita dalla dinamica a priori. Il coefficiente β (impostato a 1 nella derivazione standard, ma con schedulazione adattiva in DreamerV3) bilancia i due obiettivi.

Le statistiche sufficienti per l'RSSM sono le statistiche di x_t necessarie per minimizzare la perdita di ricostruzione: per un decoder gaussiano, queste sono la media e la varianza di x_t condizionate sullo stato latente — ovvero tutte le statistiche di secondo ordine dell'osservazione. Per un decoder categorico (come nelle varianti VQ-VAE), le statistiche sufficienti sono la distribuzione discreta dei codici. In entrambi i casi, l'RSSM è costretto a codificare tutto ciò che è necessario per la ricostruzione — una statistica sufficiente in senso Fisher per $p(x_t | h_t, z_t)$.

L'ELBO di Var-JEPA. L'ELBO di Var-JEPA, derivato nella sezione 3.3.1, è la somma di cinque termini privi di ricostruzione:

$$\mathcal{L}_{\text{VarJEPA}} = \mathbb{E}_q [-\|g_\psi(z_c) - z_t\|^2 / \sigma^2 - \log \sigma^2] - D_{\text{KL}}(q_\varphi(z_t | x_t) \parallel p(z_t)) - D_{\text{KL}}(q_\varphi(z_c | x_c) \parallel p(z_c)) + \lambda_{\text{cov}} \cdot \mathcal{L}_{\text{decorr}}(z_t, z_c) + \lambda_{\text{SIG}} \cdot \text{SIGReg}(z_t)$$

dove il primo termine è la perdita di predizione latente (non di ricostruzione), il secondo e il terzo sono le penalità KL sui posteriori, il quarto è una penalità di decorrelazione e il quinto è il regolarizzatore di gaussianità isotropa derivato nel Capitolo 17. Le statistiche sufficienti per Var-JEPA sono molto più parsimoniose: sono le statistiche di x_t necessarie per produrre z_t da cui il predittore g_ψ possa predire z_c — ovvero le statistiche di secondo ordine delle rappresentazioni latenti, non dell'osservazione grezza. Qualsiasi variazione di x_t che non altera la distribuzione di z_t è invisibile all'obiettivo: l'encoder Var-JEPA può ignorare liberamente le dimensioni di x_t irrilevanti per la predittività latente.

La relazione formale tra le due statistiche sufficienti. Sia $S_{\text{RSSM}}(x_t)$ la statistica sufficiente dell'RSSM e $S_{\text{VarJEPA}}(x_t)$ quella di Var-JEPA. La relazione tra le due è di inclusione: nel caso limite in cui il decoder p_θ dell'RSSM è un'iniezione (il decoder è biiettivo, ovvero non comprime informazione), $S_{\text{RSSM}} = x_t$ (la statistica sufficiente è l'osservazione intera).

Var-JEPA produce invece $S_VarJEPA \subseteq x_t$ nel senso dell'informazione mutua: $I(z_t; x_t)\{VarJEPA\} \leq I(z_t; x_t)\{RSSM\}$ a parità di dimensionalità dello spazio latente.

Questo ordinamento inverso si inverte nella direzione causalmente rilevante: $I(z_t; a_t \rightarrow x_{t+1})\{VarJEPA\} \geq I(z_t; a_t \rightarrow x_{t+1})\{RSSM\}$, dove $a_t \rightarrow x_{t+1}$ indica l'effetto causale dell'azione sulla transizione. L'argomento è che l'obiettivo di predizione latente di Var-JEPA seleziona esattamente la componente di z_t predittivamente rilevante per la transizione, mentre l'obiettivo di ricostruzione dell'RSSM alloca capacità rappresentativa anche alle componenti di x_t irrilevanti per la transizione (la tessitura del fondo, il rumore di illuminazione). Formally, se ϵ_t rappresenta la componente di x_t che non porta informazione sulla transizione futura:

$$I(S_RSSM(x_t); \epsilon_t) = I(x_t; \epsilon_t) = H(\epsilon_t) > 0$$

$$I(S_VarJEPA(x_t); \epsilon_t) \approx 0$$

Le rappresentazioni di Var-JEPA sono informativamente più efficienti nel senso che concentrano la capacità latente disponibile sull'informazione predittivamente rilevante, ignorando il rumore irrilevante. Questo è il meccanismo formale per cui si attende che i World Models Var-JEPA generalizzino meglio a nuovi ambienti: le loro rappresentazioni non sono contaminate dall'informazione specifica del dominio di training.

Le condizioni di equivalenza. Le due famiglie convergono verso lo stesso spazio di soluzioni sotto tre condizioni simultanee:

(E1) Il decoder $p_\theta(x_t | h_t, z_t)$ dell'RSSM ha dimensione nulla dello spazio nullo: il decoder è biiettivo (nessuna compressione di informazione). (E2) La componente ϵ_t irrilevante per la transizione è ortogonale alla componente causalmente rilevante nello spazio delle osservazioni: $Cov(\epsilon_t, x_t - \epsilon_t) = 0$. (E3) La capacità dello spazio latente è sufficiente a codificare separatamente l'informazione rilevante e quella irrilevante: $\dim(z_t) \geq r_eff + \dim(\epsilon_t)$.

Quando (E1)-(E3) sono soddisfatte, l'RSSM può in linea di principio apprendere una soluzione che alloca le prime r_eff dimensioni all'informazione predittivamente rilevante e le rimanenti al rumore ϵ_t — raggiungendo lo stesso comportamento di Var-JEPA. Ma questa allocazione ottimale non è garantita dall'obiettivo dell'RSSM: emerge soltanto se l'ottimizzazione converge verso il punto corretto, il che è difficile in presenza di alte dimensionalità e landscape non-convesso. Var-JEPA, al contrario, non può allocare capacità al rumore ϵ_t per costruzione: non esiste un obiettivo di ricostruzione che ricompensi la codifica di ϵ_t nelle rappresentazioni.

Ne consegue che, a parità di capacità dello spazio latente, Var-JEPA usa la capacità disponibile in modo più efficiente di RSSM nel regime di sample efficiency — quando i dati di interazione con l'ambiente sono scarsi e ogni bit

di informazione rilevante conta. Nel regime di grande scala (molti dati, grande capacità latente), le due famiglie dovrebbero convergere a prestazioni simili perché la distinzione tra informazione rilevante e irrilevante diventa asintoticamente meno critica. La domanda empirica rilevante per il confronto DreamerV3 vs Var-JEPA è quindi precisamente il regime di scarsità dei dati di interazione: nei benchmark come DMControl con 100.000 passi di interazione, Var-JEPA dovrebbe avere un vantaggio sistematico; con 10.000.000 di passi, il vantaggio dovrebbe ridursi o scomparire. Questo è il tipo di ablation che il sotto-programma IX del Capitolo 20 propone di eseguire sistematicamente.

11.5 Il teorema del gradiente di policy per le opzioni nello spazio latente JEPA: derivazione completa

La formalizzazione del Reinforcement Learning gerarchico attraverso il framework delle opzioni, introdotta nella sezione 11.1, lascia aperta una questione cruciale: come si calcola il gradiente dell'obiettivo rispetto ai parametri delle policy che operano ai diversi livelli della gerarchia? La risposta richiede una versione del Policy Gradient Theorem — il risultato fondamentale di Sutton, McAllester, Singh e Mansour (2000) che connette il gradiente della performance alle derivate parziali delle policy logaritmiche — estesa alla struttura semi-Markoviana delle opzioni e, nel caso specifico del paradigma JEPA, all'operazione nello spazio latente $\mathbb{Z} = \mathbb{R}^d$ prodotto dall'encoder.

Il teorema del gradiente per le opzioni nella sua formulazione più generale, sviluppata da Bacon, Harb e Precup nell'architettura Option-Critic (2017), distingue tre componenti del gradiente complessivo che corrispondono ai tre oggetti ottimizzabili della gerarchia: la policy intra-opzione $\pi_{\omega}(a \mid z; \theta_{\omega})$, la funzione di terminazione $\beta_{\omega}(z; \chi_{\omega})$, e la meta-policy $\pi_{\text{high}}(\omega \mid z; \psi)$. Ciascuna di queste componenti è una rete neurale parametrica che prende come input la rappresentazione latente z prodotta dall'encoder JEPA congelato e restituisce, rispettivamente, una distribuzione sulle azioni primitive, una probabilità di terminazione, e una distribuzione sulle opzioni disponibili.

Prima di enunciare il teorema, occorre precisare la notazione per le quantità di valore nella struttura a opzioni. La funzione di valore condizionata all'opzione $Q_{\Omega}(z, \omega)$ — che misura il return atteso cumulativo quando si è nello stato latente z e si esegue l'opzione ω — è definita dall'equazione di Bellman per processi semi-Markoviani:

$$Q_{\Omega}(z, \omega) = \sum_a \pi_{\omega}(a \mid z; \theta_{\omega}) \cdot [r_{\text{low}}(z, a) + \gamma \cdot \mathbb{E}_{z' \sim P(\cdot \mid z, a)} [(1 - \beta_{\omega}(z'; \chi_{\omega})) \cdot Q_{\Omega}(z', \omega) + \beta_{\omega}(z'; \chi_{\omega}) \cdot V_{\Omega}(z')]]$$

dove $r_{\text{low}}(z, a) = r_{\text{ext}}(z, a) + \lambda_{\text{intr}} \cdot r_{\text{intr}}(z, a)$ è la funzione di ricompensa composita introdotta nella sezione 11.2, e $V_{\Omega}(z) = \sum_{\omega} \pi_{\text{high}}(\omega \mid z; \psi) \cdot Q_{\Omega}(z, \omega)$ è la funzione di valore della meta-policy. La funzione di valore

condizionata all'azione e all'opzione $Q_U(z, \omega, a)$ — chiamata "upon-arrival value function" nella terminologia di Bacon et al. — richiede una definizione separata, poiché condiziona sia sull'opzione selezionata sia sull'azione compiuta:

$$Q_U(z, \omega, a) = r_{\text{low}}(z, a) + \gamma \cdot \mathbb{E}_{\{z'\}} [(1 - \beta_{\omega}(z')) \cdot Q_{\Omega}(z', \omega) + \beta_{\omega}(z') \cdot V_{\Omega}(z')]$$

La distinzione tra Q_{Ω} e Q_U è architetturealmente significativa: Q_{Ω} misura il valore di un'opzione al momento della sua selezione (prima che l'azione corrente sia scelta secondo π_{ω}), mentre Q_U misura il valore dopo che l'azione è stata eseguita ma prima di osservare la transizione di stato. Il vantaggio dell'opzione $A_{\Omega}(z, \omega) = Q_{\Omega}(z, \omega) - V_{\Omega}(z)$, che sarà centrale nella derivazione del gradiente di terminazione, misura di quanto sia vantaggioso continuare a eseguire l'opzione ω rispetto a lasciare che la meta-policy selezioni liberamente tra tutte le opzioni.

Teorema del Gradiente Intra-Opzione. Sia $J(\theta_{\omega}) = \mathbb{E}_{\{z_0 \sim \mu\}} [V_{\Omega}(z_0; \theta_{\omega}, \chi_{\omega}, \psi)]$ la performance media sotto la politica gerarchica congiunta. Allora il gradiente rispetto ai parametri θ_{ω} della policy intra-opzione π_{ω} è:

$$\partial J / \partial \theta_{\omega} = \mathbb{E}_{\{z_t, \omega_t = \omega \sim \mu_{\Omega}\}} [\Sigma_t \gamma^t \cdot \nabla_{\{\theta_{\omega}\}} \log \pi_{\omega}(a_t | z_t; \theta_{\omega}) \cdot Q_U(z_t, \omega_t, a_t)]$$

dove $\mu_{\Omega}(z, \omega)$ è la misura di occupazione congiunta (stato latente, opzione) sotto la politica gerarchica π_{high} e il processo di terminazione β_{ω} , definita come $\mu_{\Omega}(z, \omega) = \Sigma_{\{t=0\}}^{\infty} \gamma^t \cdot P(z_t = z, \omega_t = \omega | z_0 \sim \mu, \pi_{\text{high}}, \{\pi_{\omega}\}, \{\beta_{\omega}\})$.

La dimostrazione procede per analogia con la dimostrazione del Policy Gradient Theorem standard. Si considera la derivata parziale di Q_U rispetto a θ_{ω} :

$$\partial Q_U(z, \omega, a) / \partial \theta_{\omega} = \gamma \cdot \mathbb{E}_{\{z'\}} [(1 - \beta_{\omega}(z')) \cdot \partial Q_{\Omega}(z', \omega) / \partial \theta_{\omega}]$$

dove il termine con $\beta_{\omega}(z') \cdot V_{\Omega}(z')$ non contribuisce al gradiente perché — quando l'opzione termina in z' — la meta-policy seleziona una nuova opzione indipendentemente da θ_{ω} . Espandendo ricorsivamente $\partial Q_{\Omega} / \partial \theta_{\omega}$ e usando la regola della catena sul termine $\Sigma_a \pi_{\omega}(a | z; \theta_{\omega}) \cdot Q_U(z, \omega, a)$:

$$\begin{aligned} \partial Q_{\Omega}(z', \omega) / \partial \theta_{\omega} &= \Sigma_a [\partial \pi_{\omega}(a | z'; \theta_{\omega}) / \partial \theta_{\omega} \cdot Q_U(z', \omega, a) + \pi_{\omega}(a | z'; \theta_{\omega}) \cdot \partial Q_U(z', \omega, a) / \partial \theta_{\omega}] \\ &= \Sigma_a \pi_{\omega}(a | z'; \theta_{\omega}) \cdot \nabla_{\{\theta_{\omega}\}} \log \pi_{\omega}(a | z'; \theta_{\omega}) \cdot Q_U(z', \omega, a) + \Sigma_a \pi_{\omega}(a | z'; \theta_{\omega}) \cdot \partial Q_U(z', \omega, a) / \partial \theta_{\omega} \end{aligned}$$

Sciogliendo la ricorsione su un orizzonte infinito e notando che i termini di secondo ordine convergono per $|\gamma| < 1$, si ottiene esattamente la somma ponderata dalla misura di occupazione μ_{Ω} , che è la tesi.

Teorema del Gradiente di Terminazione. Il gradiente rispetto ai parametri χ_{ω} della funzione di terminazione β_{ω} è:

$$\partial J / \partial \chi_{-\omega} = - \mathbb{E}_{\{z'_{t+1}, \omega_t = \omega\} \sim \mu_{-\Omega}} [\sum_t \gamma^{t+1} \cdot \nabla \{ \chi_{-\omega} \} \beta_{-\omega}(z'_{t+1}; \chi_{-\omega}) \cdot A_{-\Omega}(z'_{t+1}, \omega_t)]$$

Il segno negativo ha un'interpretazione economica precisa: poiché $A_{-\Omega}$ misura il vantaggio di *continuare* l'opzione rispetto a terminare, il gradiente spinge $\chi_{-\omega}$ nella direzione che aumenta $\beta_{-\omega}$ (aumenta la probabilità di terminazione) quando $A_{-\Omega} < 0$ — ossia quando continuare è svantaggioso — e la riduce quando $A_{-\Omega} > 0$. In altri termini, la funzione di terminazione apprende a terminare l'opzione negli stati in cui sarebbe meglio che la meta-policy scelga liberamente, e a prolungarla negli stati in cui il contributo specifico di quell'opzione è positivo. La dimostrazione di questo risultato si ottiene differenziando J rispetto a $\chi_{-\omega}$ attraverso il termine $\beta_{-\omega}$ nella definizione di $Q_{-\Omega}$, tenendo conto del fatto che la terminazione influenza il flusso di controllo tra l'opzione corrente e la meta-policy, non la funzione di transizione di stato.

Il gradiente della meta-policy. L'aggiornamento della meta-policy $\pi_{\text{high}}(\omega | z; \psi)$ è formalmente analogo al Policy Gradient Theorem standard per spazi di azione discreti (le opzioni), con $Q_{-\Omega}$ nel ruolo della funzione di valore azione:

$$\partial J / \partial \psi = \mathbb{E}_{\{z_{t_k} \sim \mu_{\text{switch}}\}} [\sum_k \nabla \psi \log \pi_{\text{high}}(\omega_k | z_{t_k}; \psi) \cdot \hat{A}_{-\Omega}(z_{t_k}, \omega_k)]$$

dove la somma è sui momenti di *cambio opzione* t_k (i passi in cui la terminazione ha avuto luogo), μ_{switch} è la distribuzione di occupazione condizionata ai momenti di switch, e $\hat{A}_{-\Omega} = Q_{-\Omega} - b(z)$ è il vantaggio centrato con la baseline $b(z) = V_{-\Omega}(z)$, che riduce la varianza dell'estimatore del gradiente senza alterarne il valore atteso.

La struttura gerarchica del gradiente — tre teoremi distinti per le tre componenti parametriche — produce un algoritmo di training end-to-end in cui ciascuna componente si aggiorna attraverso il segnale di errore appropriato: l'errore di predizione dell'azione per $\pi_{-\omega}$, il segno del vantaggio dell'opzione per $\beta_{-\omega}$, e l'advantage complessiva dell'opzione per π_{high} . Questo contrasto con gli approcci che ottimizzano le tre componenti attraverso lo stesso loss scalare — una scelta che contamina i gradienti e rallenta la convergenza, come documentato da Bacon et al. nei loro esperimenti su Four Rooms e Montezuma's Revenge.

L'integrazione del World Model per la stima dell'advantage. La contribuzione specifica del paradigma JEPA alla struttura appena derivata risiede nella possibilità di stimare Q_U attraverso rollout immaginati nel World Model, piuttosto che richiedere campioni di interazione reale con l'ambiente. Il World Model latente $f_{-\psi}(z_t, a_t) \approx z_{t+1}$ permette di costruire la stima a n passi:

$$Q_U^{\{n\}}(z_t, \omega_t, a_t) = \sum_{k=0}^{n-1} \gamma^k \cdot r_{\text{low}}(\hat{z}_{t+k}, a_{t+k}) + \gamma^n \cdot [(1 - \beta_{\omega}(\hat{z}_{t+n})) \cdot Q_{\Omega}(\hat{z}_{t+n}, \omega_t) + \beta_{\omega}(\hat{z}_{t+n}) \cdot V_{\Omega}(\hat{z}_{t+n})]$$

dove la traiettoria immaginata è costruita come segue: $\hat{z}_{t+0} = z_t$, $\hat{z}_{t+k+1} = f_{\varphi}(\hat{z}_{t+k}, a_{t+k})$, con $a_{t+0} = a_t$ (l'azione effettivamente compiuta al tempo t , campionata da π_{ω} nell'ambiente reale) e $a_{t+k} \sim \pi_{\omega}(\cdot \mid \hat{z}_{t+k}; \theta_{\omega})$ per $k \geq 1$ (azioni generate dalla policy intra-opzione sull'immaginazione). I valori al termine del rollout $Q_{\Omega}(\hat{z}_{t+n}, \omega_t)$ e $V_{\Omega}(\hat{z}_{t+n})$ sono stimati da reti di valore parametriche, tipicamente Multilayer Perceptron che operano sullo spazio latente congelato.

L'estimatore ibrido $\hat{A}_U^{\{n\}}(z_t, \omega_t, a_t) = Q_U^{\{n\}} - V_{\Omega}(z_t)$ offre un compromesso bias-varianza regolabile attraverso n : per $n = 1$ si ha il TD(0) advantage, a bassa varianza ma alto bias (un singolo passo del World Model può essere impreciso); per $n \rightarrow \infty$ si ha il Monte Carlo advantage, a varianza infinita ma bias nullo; per valori intermedi $n \in \{5, 10, 20\}$ si ottiene il compromesso empiricamente ottimale documentato nelle applicazioni model-based di DreamerV3. La differenza rispetto a DreamerV3 è che in quel sistema i rollout immaginati si estendono fino all'orizzonte completo in ogni aggiornamento, mentre nell'architettura gerarchica proposta i rollout si interrompono al primo momento di terminazione dell'opzione — il che rende i rollout significativamente più corti (la durata media di un'opzione è tipicamente $O(10)$ step, molto meno degli $O(100)$ di un episodio completo) e riduce l'accumulo di errore del World Model.

La calcolabilità per backpropagation del gradiente $\partial J / \partial \theta_{\omega}$ attraverso i rollout immaginati richiede che f_{φ} sia differenziabile rispetto a z_t , che è garantito per qualsiasi predittore latente basato su MLP o Transformer. La condizione aggiuntiva che il gradiente non esploda attraverso n passi di rollout è gestita con gradient clipping sull'intera catena temporale e, nelle implementazioni recenti di DreamerV3, con la tecnica symlog che compressa l'intervallo dei valori di ricompensa.

11.6 La struttura dell'equazione di Bellman nel modello gerarchico con World Model JEPA: condizioni di convergenza e connessione con la pianificazione come inferenza

La derivazione del Policy Gradient Theorem nella sezione precedente ha operato implicitamente assumendo l'esistenza di funzioni di valore Q_{Ω} , Q_U , V_{Ω} ben definite e stabili durante il training. Questa assunzione non è ovvia in un sistema gerarchico con World Model, poiché le tre componenti — funzioni di valore, policy e World Model — evolvono simultaneamente durante l'addestramento, e la convergenza dell'intero sistema dipende dalla compatibilità

delle loro dinamiche. La presente sezione dedica l'attenzione a questo problema, derivando formalmente l'equazione di Bellman per la struttura gerarchica con World Model JEPA e analizzando le condizioni di convergenza.

L'operatore di Bellman per le opzioni. Sia $\mathcal{Q} = \{Q : \mathbb{Z} \times \Omega \rightarrow \mathbb{R}\}$ lo spazio delle funzioni di valore dell'opzione equipaggiato con la norma ℓ^∞ : $\|Q\|^\infty = \sup\{z \in \mathbb{Z}, \omega \in \Omega\} |Q(z, \omega)|$. Si definisce l'operatore di Bellman ottimale per le opzioni $T_{\Omega^*} : \mathcal{Q} \rightarrow \mathcal{Q}$ come:

$$(T_{\Omega^*} Q)(z, \omega) = \sum_a \pi_{\omega^*}(a | z) \cdot [r_{\text{low}}(z, a) + \gamma \cdot \mathbb{E}_{z'} \{ [(1 - \beta_{\omega^*}(z')) \cdot Q(z', \omega) + \beta_{\omega^*}(z') \cdot \max_{\omega'} \{Q(z', \omega')\}] \}]$$

dove π_{ω^*} e β_{ω^*} denotano le policy e le funzioni di terminazione greedy rispetto alla funzione di valore corrente Q . L'operatore T_{Ω^*} è una contrazione in norma ℓ^∞ con costante di contrazione γ :

$$\|T_{\Omega^*} Q - T_{\Omega^*} \tilde{Q}\|^\infty \leq \gamma \cdot \|Q - \tilde{Q}\|^\infty \quad \forall Q, \tilde{Q} \in \mathcal{Q}$$

Dimostrazione. Per qualsiasi (z, ω) fissato, si ha:

$$\begin{aligned} |(T_{\Omega^*} Q)(z, \omega) - (T_{\Omega^*} \tilde{Q})(z, \omega)| &\leq \gamma \cdot \mathbb{E}_{z'} \{ [(1 - \beta_{\omega^*}(z')) \cdot (Q(z', \omega) - \tilde{Q}(z', \omega)) + \beta_{\omega^*}(z') \cdot (\max_{\omega'} \{Q(z', \omega')\} - \max_{\omega'} \{\tilde{Q}(z', \omega')\})] \} \\ &\leq \gamma \cdot \mathbb{E}_{z'} \{ [(1 - \beta_{\omega^*}(z')) \cdot |Q(z', \omega) - \tilde{Q}(z', \omega)| + \beta_{\omega^*}(z') \cdot \max_{\omega'} |Q(z', \omega') - \tilde{Q}(z', \omega')|] \} \\ &\leq \gamma \cdot \mathbb{E}_{z'} \{ \|Q - \tilde{Q}\|^\infty \} = \gamma \cdot \|Q - \tilde{Q}\|^\infty \end{aligned}$$

dove si è usata la disuguaglianza $|\max_{\omega} f(\omega) - \max_{\omega} g(\omega)| \leq \max_{\omega} |f(\omega) - g(\omega)| \leq \|f - g\|^\infty$ e il fatto che $\beta_{\omega^*}(z') \in [0, 1]$ implica $(1 - \beta_{\omega^*}(z')) + \beta_{\omega^*}(z') = 1$. Poiché $\gamma < 1$ per costruzione, il Teorema del Punto Fisso di Banach garantisce l'esistenza di un unico punto fisso $Q^* = T_{\Omega^*} Q^*$ tale che $Q_n \rightarrow Q$ geometricamente: $\|Q_n - Q\|^\infty \leq \gamma^n \cdot \|Q_0 - Q\|^\infty$. ■

La contrattività dell'operatore di Bellman per le opzioni stabilisce che, se il World Model f_φ fosse esatto ($f_\varphi(z, a) = \mathbb{E}[z' | z, a]$ con errore nullo), la Q-learning gerarchica su tutto lo spazio latente convergerebbe al valore ottimale a partire da qualsiasi inizializzazione. La sfida pratica è che il World Model non è esatto, e il suo errore di predizione si accumula lungo i rollout.

Analisi dell'errore del World Model e bound sulla subottimalità. Sia $\epsilon_{\text{WM}} = \sup\{z, a\} \|f_\varphi(z, a) - \mathbb{E}[z' | z, a]\|_2$ l'errore massimo del World Model (in norma ℓ_2 sull'immagine del predittore). L'effetto di questo errore sulla stima della Q-function attraverso rollout a n passi può essere quantificato formalmente. Sia $Q^{(n)}_{\text{U}}$ la stima a n passi dell'equazione di Bellman e Q_{U}^* il valore ottimale vero. Si ha:

$$\|Q_{\text{U}}^{(n)} - Q_{\text{U}}^*\|^\infty \leq (\gamma^n / (1 - \gamma)) \cdot \|r_{\text{max}}\| + (\sum_{k=1}^n \gamma^k \cdot L_r \cdot \epsilon_{\text{WM}} \cdot k)$$

dove $L_r = \|\nabla_{z'} r_{\text{low}}\|^\infty$ è la costante di Lipschitz della funzione di ricompensa rispetto allo stato latente, e il secondo termine cresce quadraticamente in n a

causa dell'accumulo di errore lungo la traiettoria. Il compromesso è quindi preciso: aumentare n riduce il bias del return (il primo termine, che converge a zero per $n \rightarrow \infty$) ma aumenta l'errore dovuto all'imprecisione del World Model (il secondo termine, che cresce). Il valore ottimale di n^* , che minimizza il bound sulla subottimalità, è:

$$n^* = \operatorname{argmin}_n \{ (\gamma^n / (1 - \gamma)) \cdot \|r_{\max}\| + \gamma \cdot L_r \cdot \epsilon_{\text{WM}} \cdot n(n+1) / 2 \} \approx \sqrt{(\|r_{\max}\| / (L_r \cdot \epsilon_{\text{WM}} \cdot (1 - \gamma)))} \text{ per } \gamma \rightarrow 1$$

Questa espressione mostra che l'orizzonte di rollout ottimale è inversamente proporzionale alla radice quadrata dell'errore del World Model: sistemi con World Model più accurato (ϵ_{WM} piccolo) dovrebbero usare rollout più lunghi. Questo risultato fornisce una giustificazione formale alla pratica empirica osservata in DreamerV3 e nelle sue varianti, dove l'orizzonte di immaginazione viene ridotto nelle fasi iniziali del training (quando il World Model è impreciso) e aumentato nelle fasi mature.

Stabilità del training congiunto: il problema dell'instabilità reciproca. Il problema più sottile nel training gerarchico con World Model è la possibilità di instabilità *reciproca*: se il World Model f_φ cambia durante il training, le funzioni di valore stimate su di esso diventano imprecise, il che porta ad aggiornamenti scorretti delle policy, che a loro volta raccolgono dati in distribuzioni diverse da quelle usate per addestrare il World Model, amplificando l'errore. Questo ciclo di feedback può portare a un collasso del training.

La separazione architetturale del paradigma JEPA — encoder E_{JEPA} preaddestrato e congelato, World Model f_φ addestrato su rappresentazioni fisse — spezza uno degli anelli di questo ciclo. Se l'encoder è congelato, la distribuzione delle rappresentazioni latenti è fissa, e il World Model è addestrato su una distribuzione stazionaria. Il problema dell'instabilità si riduce alla covarianza del World Model con le funzioni di valore, che è gestibile con tecniche standard: target networks per le funzioni di valore (aggiornamenti EMA con tasso $\tau \ll 1$), replay buffer per la distribuzione di addestramento del World Model, e clipping del gradiente per prevenire aggiornamenti catastrofici.

Formalmente, sia $\theta = (\varphi, \{\theta_\omega\}, \{\chi_\omega\}, \psi, \varphi_V)$ il vettore di tutti i parametri ottimizzabili. Il training congiunto attraverso gradient descent con step size η ha proprietà di convergenza garantite se la matrice jacobiana del sistema di aggiornamento $J(\theta)$ soddisfa $\varrho(I - \eta \cdot J(\theta)) < 1$, dove ϱ è il raggio spettrale. Per la struttura gerarchica con encoder congelato, questa condizione si riduce alla condizione di stabilità del sistema dinamico lineare ottenuto linearizzando attorno al punto fisso, che può essere verificata numericamente per ogni architettura specifica.

Connessione con la pianificazione come inferenza variazionale. La sezione 9.2 ha mostrato come la pianificazione possa essere riformulata come inferenza variazionale in un modello generativo con variabili di controllo latenti. Questa connessione si estende naturalmente alla struttura gerarchica: le opzioni

$\omega_1, \dots, \omega_K$ sono le variabili latenti discrete al livello alto della gerarchia, mentre le azioni $a_{1:T}$ sono le variabili latenti continue al livello basso. Il modello generativo congiunto è:

$$p(o_{1:T}, a_{1:T}, \omega_{1:K}) = p(\omega_1) \cdot [\prod_k p(\omega_{k+1} \mid \omega_k, z_{t_k}) \cdot p_{\text{switch}}(t_k \mid \omega_k, z_{t_{k-1}})] \cdot [\prod_t p(a_t \mid z_t, \omega_{\kappa(t)}) \cdot p(o_{t+1} \mid z_t, a_t)]$$

dove $\kappa(t)$ è la funzione che mappa ogni passo temporale t al segmento dell'opzione k in corso, p_{switch} è la distribuzione del tempo di commutazione determinata da β_ω , e $p(a_t \mid z_t, \omega_{\kappa(t)})$ è la distribuzione delle azioni intra-opzione. La pianificazione ottimale corrisponde al problema di inferenza:

$$p^*(\omega_{1:K}, a_{1:T} \mid o_{1:T}, r = 1) \propto p(o_{1:T}, a_{1:T}, \omega_{1:K}) \cdot \exp(\sum_t r_t)$$

dove $r = 1$ è la variabile binaria che indica il raggiungimento del goal (nel formalismo del MaxEnt RL, il reward è trattato come log-likelihood di un evento ottimale). L'inferenza variazionale su questo modello — approssimando la posterior con una distribuzione fattorizzata $q(\omega_{1:K}) \cdot q(a_{1:T})$ — produce l'algoritmo di Expectation-Maximization gerarchico che alterna tra la stima delle opzioni (E-step: inferire $\omega_{1:K}$ dati $a_{1:T}$) e l'ottimizzazione delle policy (M-step: aggiornare $\{\pi_\omega\}$, $\{\beta_\omega\}$, π_{high} per massimizzare il lower bound variazionale).

Questa riformulazione evidenzia perché la struttura del reward intrinseco di JEPA — che premia la predittività del World Model piuttosto che la novità bruta — è architetturalmente compatibile con la pianificazione come inferenza: la predittività del World Model è proporzionale alla log-likelihood delle transizioni sotto il modello, che è esattamente la quantità massimizzata nel M-step. In altri termini, le opzioni che producono traiettorie altamente predicibili nel World Model JEPA corrispondono a traiettorie ad alta verosimiglianza nel modello generativo, che sono le traiettorie preferite dall'inferenza variazionale — chiudendo elegantemente il cerchio tra l'RL gerarchico basato su opzioni e la pianificazione come inferenza bayesiana.

Capitolo 12 — Probing causale negli spazi latenti JEPA

12.1 Il problema epistemologico dell'ispezione interna

Che cosa rappresentano i vettori nello spazio latente di un sistema JEPA addestrato? La domanda sembra semplice — vi si può rispondere addestrando un classificatore lineare su quelle rappresentazioni e misurando la sua

accuratezza su un task rilevante — ma questa risposta semplice confonde due questioni che la teoria dell'apprendimento tiene invece rigorosamente distinte: la domanda di cosa sia *codificato* in uno spazio latente e la domanda di cosa sia *causalmente determinante* per le previsioni prodotte da quel sistema. Distinguere queste due questioni non è un esercizio puramente filosofico. Avere un sistema che codifica informazione causalmente rilevante nei propri stati latenti è la condizione necessaria — sebbene non sufficiente — per garantire che le sue decisioni siano giustificabili, generalizzabili e modificabili in modo mirato. Questa condizione è ancora più stringente nel contesto dei sistemi JEPA, dove lo spazio latente è l'unico luogo in cui avviene la computazione rilevante: non c'è una ricostruzione in pixel-space che funga da ancoraggio alternativo.

Il programma del *probing* — la tecnica di addestrare classificatori o regressori leggeri sulle rappresentazioni latenti di una rete profonda e di usare la loro accuratezza come proxy della presenza di una certa proprietà nello spazio latente — è diventato la metodologia dominante per l'analisi interna dei modelli fondazionali. Alain e Bengio (2016) ne hanno proposto la prima formalizzazione sistematica: se uno strato di una rete codifica informazione accessibile linearmente su un task, allora si può stimare la «linearmente decodificabile information» addestrando una regressione logistica su quel strato. Il risultato di questa ispezione è un numero: l'accuratezza del probing classifier. La semplicità del metodo ne ha favorito l'adozione, ma ne ha anche oscurato i limiti profondi. Belinkov (2022) e Hewitt e Liang (2019) mostrano che l'accuratezza di un probing classifier non è un proxy affidabile della «quantità di informazione» presente nelle rappresentazioni, perché il classificatore di probing porta essa stessa un contributo computazionale non trascurabile: un classificatore lineare accurato potrebbe riflettere la propria capacità di trovare separazioni lineari piuttosto che la struttura dello spazio latente. Questa osservazione porta alla distinzione tra *probing classifiers* come strumenti di analisi esterna e come strumenti di analisi interna — distinzione che diventa critica nel momento in cui si vuole fare non soltanto analisi descrittiva, ma analisi causale.

12.2 Dalla correlazione alla causalità nel probing

Si definisca formalmente il problema. Sia $E : \mathcal{X} \rightarrow \mathbb{R}^d$ l'encoder JEPA, che mappa le osservazioni nello spazio latente d -dimensionale. Sia $\mathcal{Z} = E(\mathcal{X})$ l'immagine dell'encoder. Sia $t : \mathcal{X} \rightarrow \mathcal{T}$ una funzione che assegna a ogni osservazione un valore sul task target di interesse (per esempio, la categoria semantica dell'oggetto, la direzione della rotazione, la posizione di un patch). Un probing classifier è una funzione $g : \mathbb{R}^d \rightarrow \mathcal{T}$, scelta tipicamente nella classe delle funzioni lineari $\mathcal{G}_{\text{lin}} = \{g : g(z) = W \cdot z + b\}$, e addestrata minimizzando una perdita $\mathcal{L}_{\text{probe}}(g \circ E, t)$ su un dataset etichettato.

L'accuratezza del probing classifier misura quanto bene il task t è *linearmente decodificabile* dall'embedding $E(x)$. Questa misura è informativa: se $g \circ E$ è molto accurato, allora E codifica qualcosa di correlato con t . Ma non dice nulla sulla direzione del nesso: E potrebbe codificare t causalmente (t è una variabile

causale che influenza E), oppure t e E potrebbero avere una causa comune, oppure il nesso potrebbe essere puramente statistico — una correlazione del dataset che non riflette alcuna struttura generativa.

Il *probing causale* si distingue dal probing classico esattamente nella misura in cui si richiede di determinare non soltanto se E codifica t, ma se E *dipende causalmente* da t nel senso del do-calcolo di Pearl. La domanda formale è: per una variabile latente $z = E(x)$, la distribuzione $p(z \mid \text{do}(t = v))$ — la distribuzione interventale di z quando si impone causalmente $t = v$ — differisce da $p(z \mid t = v)$? Se le due distribuzioni coincidono per ogni valore v, allora la correlazione tra E e t non riflette dipendenza causale di E da t: entrambe potrebbero essere determinate da una causa comune non misurata. Se invece $p(z \mid \text{do}(t = v)) \neq p(z \mid t = v)$, allora E è causalmente discendente da t, e la codifica in E di informazione su t è genuinamente causale.

La distanza tra queste due distribuzioni può essere quantificata attraverso la differenza:

$$\Delta_{\text{causal}}(t, E) = \mathbb{E}_{\{v \sim p(t)\}} [D_{\text{KL}}(p(E(x) \mid \text{do}(t = v)) \parallel p(E(x) \mid t = v))]]$$

dove D_{KL} è la divergenza di Kullback-Leibler. Se $\Delta_{\text{causal}}(t, E) = 0$, la variabile t non è causalmente determinante per E anche se la correlazione osservazionale tra E e t è positiva. Se $\Delta_{\text{causal}}(t, E) > 0$, t ha un effetto causale su E proporzionale a questa quantità.

Il problema pratico è che la distribuzione interventale $p(E(x) \mid \text{do}(t = v))$ non è in generale osservabile direttamente: richiede di intervenire causalmente sulla variabile t, il che è possibile soltanto in ambienti di valutazione che ammettono manipolazioni controllate. Nel contesto dei dataset di valutazione per i sistemi JEPa — come 3DIEBench per seq-JEPa — questa condizione è soddisfatta in modo naturale: il dataset è generato mediante un processo di rendering 3D in cui le variabili di fattore (rotazione, colore del pavimento, intensità della luce) sono controllate sperimentalmente. La distribuzione $p(E(x) \mid \text{do}(t = v))$ è approssimabile dalla distribuzione empirica $p(E(x) \mid t = v)$ nel dataset, perché il processo generativo è noto e le variabili di fattore sono indipendenti per costruzione.

12.3 Il test di probing causale via do-calcolo su 3DIEBench

Il benchmark 3DIEBench, introdotto da Garrido et al. (2023) come ambiente di valutazione per il carico di equivarianza, offre le condizioni ideali per un test di probing causale. Gli oggetti 3D sono resi con variazioni controllate in tre fattori: la rotazione $SO(3)$ dell'oggetto, la tonalità del pavimento e l'intensità luminosa. Il processo generativo è fattorizzato per costruzione — le tre variabili sono campionate indipendentemente — il che significa che intervenire su una variabile (cambiare la rotazione tenendo fissi colore e luce) produce la distribuzione interventale esatta, non una distribuzione condizionale contaminata da correlazioni spurie.

Nel framework seq-JEPA, si distinguono due spazi latenti: le rappresentazioni dell'encoder $z_i = E(x_i)$ per ciascuna vista, e la rappresentazione aggregata $z_{AGG} = g((z_1, a_1), (z_2, a_2), \dots, (z_M))$ prodotta dal transformer aggregatore. Il comportamento di questi due spazi latenti rispetto alla rotazione è misurabile attraverso due metriche complementari, entrambe derivabili dalla framework del probing:

La metrica di equivarianza misura quanto bene la trasformazione relativa $a_{\{i,j\}}$ tra due viste è decodificabile dalla differenza $z_i - z_j$ (o da una funzione più complessa di (z_i, z_j)). Formalmente, si addestra un regressore $g_{eq} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^4$ a predire il quaternionione di rotazione relativa $a_{\{i,j\}}$ dalla coppia (z_i, z_j) , e si misura il coefficiente di determinazione R^2 sulla distribuzione di test. Nella notazione del do-calcolo, questo corrisponde a misurare $p(g_{eq}(z_i, z_j) \approx a_{\{i,j\}} \mid \text{do}(a_{\{i,j\}} = q))$ per quaternioni q campionati dalla distribuzione di test.

La metrica di invarianza misura quanto bene la categoria dell'oggetto è decodificabile da z_{AGG} attraverso un classificatore lineare, in modo robusto alle variazioni di rotazione. Nella notazione interventale, si verifica che $p(g_{inv}(z_{AGG}) = \text{classe} \mid \text{do}(\text{rotazione} = R))$ sia costante al variare di R — l'invarianza causale alla rotazione significa che intervenire sulla rotazione non modifica la distribuzione del classificatore di classe.

I risultati riportati da Ghaemi et al. (2025) per seq-JEPA (3, 5) — tre viste durante il training, cinque durante la validazione — mostrano un $R^2 = 0.71$ per la predizione della rotazione relativa e un'accuratezza di classificazione lineare dell'87.41%. Questo profilo è il più bilanciato tra tutti i metodi confrontati nel benchmark. SIE e ContextSSL raggiungono R^2 di equivarianza più alti (0.58 e 0.74 rispettivamente), ma con accurtezze di classificazione sensibilmente inferiori (77.49% e 80.40%). BYOL e SimCLR raggiungono accurtezze di classificazione più alte (82.90% e 81.13%) con R^2 di equivarianza molto bassi (0.12 e 0.35). La separazione architettónica introdotta da seq-JEPA — encoder equivariante, aggregatore invariante — produce quindi uno spazio latente biforcuto in cui le due proprietà coesistono senza compromessi reciproci.

Questo risultato, visto attraverso la lente del probing causale, ha un'interpretazione precisa: le rappresentazioni dell'encoder $E(x_i)$ contengono informazione causalmente determinata dalla rotazione ($\Delta_{\text{causal}}(\text{rotazione}, E) > 0$), mentre la rappresentazione z_{AGG} è causalmente invariante alla rotazione ($\Delta_{\text{causal}}(\text{rotazione}, g \circ E_{agg}) \approx 0$). La struttura architetturale del transformer aggregatore — che combina le viste (z_i, a_i) attraverso attenzione multi-testa prima di produrre z_{AGG} — funge da integrazione causale multipla delle trasformazioni, portando all'emergenza di un'invarianza che non era imposta esplicitamente come objective ma che emerge dalle proprietà statistiche del processo di aggregazione.

12.4 La condizione di identificabilità del probing causale

Perché il probing causale sia informativo — e non soltanto una misura della correlazione con le variabili di fattore del dataset — devono essere soddisfatte tre condizioni di identificabilità che generalizzano le condizioni di Khemakhem et al. (2020) per l'IVAE al contesto del probing.

La prima condizione è la *sufficienza contestuale*: le variabili di fattore latenti z che si vogliono testare devono essere causalmente complete rispetto all'osservazione, nel senso che $p(x | z)$ non dipende da nessun'altra variabile causale non osservata. Nel contesto di 3DIEBench questa condizione è soddisfatta per costruzione: le tre variabili di fattore specificano completamente il processo di rendering. In dataset naturali — immagini del mondo reale — questa condizione è inevitabilmente violata, il che significa che i risultati del probing causale su dataset naturali non possono mai essere interpretati come prove definitive di dipendenza causale.

La seconda condizione è l'*indipendenza interventale*: le variabili di fattore devono essere manipolabili indipendentemente l'una dall'altra, cioè intervenire su t_1 non deve alterare la distribuzione di t_2 . Questa condizione corrisponde all'assunzione di modularità degli SCM di Pearl: ciascun meccanismo causale è autonomo rispetto agli interventi sugli altri meccanismi. Anche questa condizione è soddisfatta per costruzione in 3DIEBench ma non in dataset naturali, dove le variabili di fattore sono tipicamente correlate (oggetti rotondi tendono a rotolare, oggetti grandi tendono a essere più pesanti, ecc.).

La terza condizione è la *distinguibilità delle direzioni causali*: deve essere possibile distinguere la direzione $t \rightarrow E$ dalla direzione $E \rightarrow t$. In un sistema JEPA addestrato con supervisione causale questa distinzione è garantita dall'architettura stessa: le osservazioni x sono fornite all'encoder, le etichette t non entrano nel processo generativo in fase di inferenza (soltanto durante il training con action conditioning). Il flusso causale va quindi necessariamente da t (che determina il processo generativo delle osservazioni x) ad $E(x)$ (che elabora le osservazioni), non viceversa.

Quando le tre condizioni sono soddisfatte, il coefficiente R^2 del probing causale è un proxy valido del grado di dipendenza causale tra il fattore t e la rappresentazione E . La formula di decomposizione della varianza si scrive:

$$R^2_{\text{causal}}(t, E) = 1 - \text{Var}[t - g_{\text{best}}(E(x))] / \text{Var}[t]$$

dove g_{best} è il miglior predittore lineare di t a partire da $E(x)$ nella classe dei predittori stimati su interventi controllati $\text{do}(t = v)$. Questa formula è identica all' R^2 standard, ma è computata su dati interventali piuttosto che osservazionali. Nei dataset con design sperimentale fattoriale come 3DIEBench, i due valori coincidono, il che giustifica l'uso diretto dell' R^2 come metrica di equivarianza causale.

12.4.1 La geometria della sufficienza lineare: quando il probing classico è fedele al probing causale

La sezione 12.4 ha enunciato tre condizioni di identificabilità necessarie perché il probing causale sia informativo, e ha mostrato come il coefficiente R^2 — calcolato su dati interventali in un benchmark a design fattoriale come 3DIEBench — costituisca una misura valida della dipendenza causale. Rimane tuttavia senza risposta una questione più sottile e di rilievo architetturale: in quali circostanze il probing *lineare* — ossia il probing basato su classificatori o regressori della forma $g(z) = Wz + b$ — è sufficiente a misurare la dipendenza causale con piena fedeltà? Quando, invece, un probing non lineare recupererebbe informazione causale che il probing lineare non vede?

La risposta richiede di analizzare la geometria delle distribuzioni condizionali $p(z \mid \text{do}(t = v))$ nello spazio latente. Si consideri la versione scalare del problema: $t \in \mathbb{R}$ è una variabile di fattore continua (per esempio l'angolo di rotazione in 3DIEBench, misurato in radianti), e $z = E(x) \in \mathbb{R}^d$ è il vettore di embedding JEPA. Il regressore lineare ottimale — nel senso di minima varianza tra tutti i regressori lineari non distorti — è $g^*(z) = w^T z + b$ con:

$$w^* = \text{Cov}(z, t) / \text{Var}(t) \cdot (\text{Cov}(z, z^T))^{-1} \cdot \text{Cov}(z, t) \text{ in forma condensata,}$$

oppure, nella notazione dell'equazione normale: $w^* = \Sigma_z^{-1} \cdot \sigma_{\{zt\}}$, dove $\Sigma_z = \text{Cov}(z, z^T)$ e $\sigma_{\{zt\}} = \text{Cov}(z, t)$. Il coefficiente di determinazione corrispondente è:

$$R^2_{\text{lin}}(t, E) = \sigma_{\{zt\}}^T \cdot \Sigma_z^{-1} \cdot \sigma_{\{zt\}} / \text{Var}(t)$$

Parallelamente, il regressore bayesianamente ottimale — che minimizza il MSE tra tutte le funzioni misurabili di z — è $g_{\text{NL}}^*(z) = \mathbb{E}[t \mid z]$, la speranza condizionale. La differenza tra le due qualità di previsione è la *lacuna di linearità*:

$$\Delta_{\text{lin}}(t, E) = R^2_{\text{NL}}(t, E) - R^2_{\text{lin}}(t, E) = 1 - \text{MSE}(g_{\text{NL}}^*) / \text{Var}(t) - R^2_{\text{lin}}(t, E)$$

dove $\text{MSE}(g_{\text{NL}}^*) = \text{Var}(t) - \text{Var}(\mathbb{E}[t \mid z])$ per la decomposizione della varianza totale.

Teorema di sufficienza lineare. Il probing lineare è sufficiente — ossia $\Delta_{\text{lin}} = 0$ e $R^2_{\text{lin}} = R^2_{\text{NL}}$ — se e soltanto se la speranza condizionale $\mathbb{E}[t \mid z]$ è una funzione affine di z .

Dimostrazione. La direzione "se": se $\mathbb{E}[t \mid z] = w^T z + b$, allora il regressore ottimale è lineare, quindi $g_{\text{NL}}^* = g^*$ e i due R^2 coincidono. La direzione "soltanto se": se $\Delta_{\text{lin}} > 0$, allora $\mathbb{E}[t \mid z] \neq w^T z + b$ per nessuna scelta di w e b

(perché altrimenti g_{NL}^* sarebbe lineare e il probing lineare sarebbe ottimale).

■

Questa condizione può essere riformulata in termini delle distribuzioni condizionali attraverso il teorema di Bayes: $\mathbb{E}[t | z]$ è affine in z se e soltanto se $p(z | t)$ appartiene alla famiglia esponenziale con sufficiente statistica lineare in z e parametri lineari in t . In particolare, se $p(z | t) = \mathcal{M}(\mu(t), \Sigma)$ con Σ costante (omoschedasticità) e $\mu(t) = At + b$ (embedding affine del fattore), allora $\mathbb{E}[t | z] = \mu(t)^\top \Sigma^{-1} z / \|\mu\|^2_{\Sigma^{-1}}$ che è lineare in z .

La connessione con l'iVAE e la curvilinearità del codice. Il risultato del Capitolo 17.4.1 stabilisce che, sotto le condizioni iVAE, il fattore latente vero $s^{\{i\}}$ è recuperabile dallo spazio JEPA attraverso $z^{\{i\}} = \varphi_i(s^{\{i\}})$ dove φ_i è una funzione monotona. Se φ_i è affine ($\varphi_i(s) = a_i \cdot s + b_i$), allora la relazione tra $s^{\{i\}}$ e $z^{\{i\}}$ è lineare, il probing lineare è sufficiente, e $\Delta_{lin} = 0$. Se φ_i è nonlineare — per esempio quadratica, logaritmica, o a forma di sigmoide — il probing lineare sottostima la dipendenza causale nella misura della curvilinearità di φ_i .

La quantità che misura questa curvilinearità è il **momento di non-linearità di φ_i** :

$$\eta(\varphi_i) = \mathbb{E}_{s^{\{i\}}} [(\varphi_i(s^{\{i\}}) - \alpha_i \cdot s^{\{i\}} - \beta_i)^2] / \text{Var}(s^{\{i\}})$$

dove $\alpha_i = \text{Cov}(\varphi_i(s), s) / \text{Var}(s)$ e $\beta_i = \mathbb{E}[\varphi_i(s)] - \alpha_i \cdot \mathbb{E}[s]$ sono la proiezione lineare migliore di φ_i . Si verifica che $\Delta_{lin} = \eta(\varphi_i) \cdot (1 - \eta(\varphi_i)) / (1 - \eta(\varphi_i) + R^2_{causal}(s^{\{i\}}, E))$ a meno di termini del secondo ordine, una formula che esprime la lacuna di linearità come funzione della curvilinearità del codice e della forza della dipendenza causale. In particolare, per codifiche molto lineari ($\eta \rightarrow 0$) la lacuna è piccola indipendentemente da R^2_{causal} , mentre per codifiche altamente nonlineari ($\eta \rightarrow 1$) la lacuna può essere grande anche quando la dipendenza causale è forte.

Questo risultato ha un'implicazione architettonica diretta per la valutazione dei sistemi JEPA: quando il probing lineare produce R^2 basso su un fattore che ci si aspetta sia causalmente rilevante (per esempio la posizione dell'oggetto in un task di grasping), bisogna verificare se la bassa accuratezza riflette assenza di dipendenza causale ($\Delta_{causal} \approx 0$) oppure curvilinearità del codice ($\Delta_{lin} > 0$). I due scenari hanno implicazioni architetturelle molto diverse: il primo suggerisce che il fattore non è codificato, il secondo che è codificato in modo non lineare e un probing con rete MLP superficiale recupererebbe la dipendenza perduta.

Quantificazione della curvilinearità via probing incrementale. Una procedura operativa per stimare $\eta(\varphi_i)$ senza conoscere φ_i in forma chiusa consiste nel confrontare le performance di probing con regressori a capacità crescente. Si addestrano sulla stessa rappresentazione JEPA congelata tre classi di regressori: lineare (L), MLP a un layer nascosto con k unità (MLP_k) con k

$\in \{16, 64, 256\}$, e kernel ridge regression con kernel polinomiale di grado 2 (KR2). Se $R^2_{\text{lin}} \approx R^2_{\{\text{MLP}_k\}} \approx R^2_{\{\text{KR2}\}}$, la codifica è approssimativamente lineare e Δ_{lin} è piccola. Se $R^2_{\{\text{KR2}\}} \gg R^2_{\text{lin}}$ con $R^2_{\{\text{MLP}_k\}}$ che cresce monotonicamente in k , la codifica è nonlineare e la lacuna è quantificata da $R^2_{\{\text{KR2}\}} - R^2_{\text{lin}}$. Questo protocollo di *probing incrementale*, che è una specializzazione della valutazione standard della curva bias-varianza, costituisce una diagnostica strutturata per l'analisi della geometria delle rappresentazioni JEPA nei task di controllo e navigazione.

12.7.1 Il test statistico della confusione causale: struttura formale e applicazione agli spazi JEPA

La sezione 12.7 ha identificato il problema delle correlazioni spurie come il limite fondamentale del probing causale in dataset non sperimentali, e ha proposto il test di robustezza causale (valutazione su distribuzioni OOD) come strategia principale per separare dipendenza causale genuina da correlazione contingente. Una formalizzazione statistica di questo test — che ne precisi la potenza, il livello di significatività, e le condizioni di applicabilità — manca nella trattazione precedente ed è l'oggetto della presente sottosezione.

Il problema statistico della confusione causale nel probing. Sia dato un dataset $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^N$ generato da un processo ignoto che può essere di due tipi: (a) il tipo causale, in cui $p(z \mid t = v) = p(z \mid \text{do}(t = v))$ per ogni v (nessuna confusione, equivalenza tra distribuzione osservazionale e interventale), oppure (b) il tipo confuso, in cui $p(z \mid t = v) \neq p(z \mid \text{do}(t = v))$ per almeno qualche valore v (la correlazione tra z e t è parzialmente o totalmente spurea). L'obiettivo è decidere tra le due ipotesi usando soltanto i dati di $\mathcal{D}$.

Quando il dataset ha un design fattoriale sperimentale — come 3DIEBench, dove le variabili di fattore sono controllate dal processo generativo — la distinzione è banale: per costruzione, l'ipotesi (a) è vera. Il test diventa non banale quando si vuole verificare se le proprietà misurate su 3DIEBench si trasferiscono a dataset naturali, dove la struttura causale è sconosciuta.

Il test di Neyman-Pearson per la confusione causale. Sia $p_0(z \mid t) = p(z \mid \text{do}(t))$ la distribuzione interventale stimata nel dataset sperimentale (tipo a), e $p_1(z \mid t) = p(z \mid t)$ la distribuzione osservazionale stimata nel dataset naturale (tipo b). L'ipotesi nulla è $H_0: p_0 = p_1$ (nessuna confusione), e l'alternativa è $H_1: p_0 \neq p_1$ (confusione causale presente). Per il Lemma di Neyman-Pearson, il test più potente per questa coppia di ipotesi semplici è il test del rapporto di verosimiglianza:

$$\Lambda(\mathcal{D}) = \prod_{i=1}^N p_1(z_i \mid t_i) / p_0(z_i \mid t_i)$$

con regione critica $\Lambda > \tau$ per una soglia τ che controlla il livello di significatività $\alpha = P(\Lambda > \tau \mid H_0)$.

Nella pratica, la distribuzione di p_0 e p_1 non è nota in forma chiusa. Se si approssima $p(z \mid t = v)$ con una Gaussiana $\mathcal{N}(\mu(v), \Sigma(v))$ stimata empiricamente nei due dataset (sperimentale e naturale), il logaritmo del rapporto di verosimiglianza si semplifica in:

$$\log \Lambda = \sum_i [-(z_i - \mu_1(t_i))^T \Sigma_1(t_i)^{-1} (z_i - \mu_1(t_i))/2 + (z_i - \mu_0(t_i))^T \Sigma_0(t_i)^{-1} (z_i - \mu_0(t_i))/2 + (\log |\Sigma_0(t_i)| - \log |\Sigma_1(t_i)|)/2]$$

Sotto H_0 ($\mu_0 = \mu_1$, $\Sigma_0 = \Sigma_1$ per ogni v), il rapporto di verosimiglianza è unitario e $\log \Lambda = 0$. Sotto H_1 , la divergenza KL media tra le due distribuzioni è:

$$\mathbb{E}_t[D_{KL}(p_1(z|t) \parallel p_0(z|t))] = \Sigma_{\text{causal}}(t, E) / 2 + \text{termini di ordine superiore}$$

dove Σ_{causal} è una funzione della differenza tra i parametri μ_1 , Σ_1 e μ_0 , Σ_0 . Per sample size sufficientemente grande, il test ha potenza $1 - o(1)$ contro qualsiasi H_1 con $D_{KL} > \epsilon$, con ϵ decrescente in N .

La statistica MMD come alternativa non parametrica. Quando la Gaussianità di $p(z \mid t)$ non è un'approssimazione adeguata — che si verifica tipicamente in spazi latenti profondi con distribuzioni multimodali — il test del rapporto di verosimiglianza gaussiano perde potenza. Un'alternativa non parametrica che non richiede ipotesi distributive è il test basato sul Maximum Mean Discrepancy (MMD), già incontrato nella sezione 17.3.2 come forma kernel di SIGReg.

Per il test di confusione causale, il MMD tra la distribuzione osservazionale e interventale del fattore condizionato è:

$$\text{MMD}^2(p_{\text{obs}}, p_{\text{int}}; k) = \mathbb{E}_{\{z, z' \sim p_{\text{obs}}\}}[k(z, z')] - 2 \cdot \mathbb{E}_{\{z \sim p_{\text{obs}}, \hat{z} \sim p_{\text{int}}\}}[k(z, \hat{z})] + \mathbb{E}_{\{\hat{z}, \hat{z}' \sim p_{\text{int}}\}}[k(\hat{z}, \hat{z}')]]$$

dove k è il kernel gaussiano e le distribuzioni $p_{\text{obs}} = p(z \mid t = v_{\text{obs}})$ e $p_{\text{int}} = p(z \mid \text{do}(t = v_{\text{int}}))$ sono stimate empiricamente. Sotto H_0 (assenza di confusione), MMD^2 ha speranza zero e il test rifiuta H_0 quando MMD^2 supera una soglia critica calcolata per permutazione. Sotto H_1 , l'aspettazione di MMD^2 è $\mathbb{E}[\text{MMD}^2] = D^2_{\text{MMD}}(p_{\text{obs}}, p_{\text{int}}) > 0$, e il test è consistente.

Il test MMD si inserisce naturalmente nell'infrastruttura computazionale di SIGReg: gli stessi kernel gaussiani e le stesse proiezioni casuali usate per regolarizzare il training possono essere riutilizzati a valutazione per stimare MMD^2 tra dataset sperimentale e naturale, rendendo il test di confusione causale un'estensione a costo marginale dell'infrastruttura già presente.

La decomposizione della varianza del probing R^2 in componenti causale e spuria. Il test della confusione causale permette di derivare una decomposizione del coefficiente R^2 del probing in tre componenti separate: la componente causale autentica R^2_{causal} (la varianza di t spiegata dalla codifica causale genuina), la componente spuria R^2_{spurious} (la varianza spiegata da correlazioni contingenti del dataset), e la componente residua $1 - R^2_{\text{causal}} - R^2_{\text{spurious}}$ (la varianza non spiegata da nessuna delle due).

Formalmente, sia $\hat{e}_{\text{causal}} = g^*(z_{\text{int}})$ il residuo del probing lineare sul dataset interventale e $\hat{e}_{\text{obs}} = g^*(z_{\text{obs}})$ il residuo sul dataset osservazionale. Allora:

$$R^2_{\text{obs}} = 1 - \text{Var}(\hat{e}_{\text{obs}}) / \text{Var}(t) \quad R^2_{\text{causal}} = 1 - \text{Var}(\hat{e}_{\text{causal}}) / \text{Var}(t) \\ R^2_{\text{spurious}} \approx R^2_{\text{obs}} - R^2_{\text{causal}}$$

Se $R^2_{\text{obs}} \gg R^2_{\text{causal}}$, la maggior parte della correlazione tra z e t nel dataset osservazionale è di natura spuria. Se $R^2_{\text{obs}} \approx R^2_{\text{causal}}$, la correlazione è causalmente autentica. La differenza $R^2_{\text{spurious}} = R^2_{\text{obs}} - R^2_{\text{causal}}$ è l'*indice di confusione causale* del dataset: un valore alto indica che il dataset naturale ha correlazioni sistematiche tra la variabile di fattore e le osservazioni che non riflettono il meccanismo causale genuino, e che i risultati del probing su quel dataset non sono trasferibili ad ambienti con distribuzioni di fattori diverse.

Nel contesto della valutazione dei sistemi JEPA destinati al deployment in ambienti fisici, questo indice fornisce una misura operativa della fragilità delle rappresentazioni rispetto alle distribuzioni fuori campione: sistemi con R^2_{spurious} alto tendono a degradare maggiormente in ambienti con distribuzioni di fattori diverse da quella di training, esattamente perché parte della loro "apparente competenza" nel probing osservazionale riflette correlazioni contingenti del dataset piuttosto che strutture causali robuste.

12.5 Probing causale e V-JEPA 2.1: il ruolo del context loss

V-JEPA 2.1 introduce una modifica architetturale fondamentale rispetto a V-JEPA 2: la context loss $\mathcal{L}_{\text{ctx}}$, che supervisa non soltanto i token mascherati ma anche i token contestuali visibili. Nella notazione del paper di Mur-Labadia et al. (2026), questa perdita è:

$$\mathcal{L}_{\text{context}} = (1/|C|) \sum_{\{i \in C\}} \lambda_i \cdot \|P_{\phi}(E_{\theta}(x), \Delta_y)_i - \text{sg}(E_{\theta}(y)_i)\|_1$$

dove C è l'insieme degli indici dei token contestuali, Δ_y è il set di token-maschera che specificano le posizioni spazio-temporali dei patch mascherati, E_{θ} è l'encoder target (exponential moving average di E_{θ}), e il peso λ_i è una funzione inversa della radice quadrata della distanza minima tra il token i e il token mascherato più vicino: $\lambda_i = \lambda / \sqrt{d_{\text{min}}(i, M)}$. Questa ponderazione favorisce i patch contestuali adiacenti alle regioni mascherate, imponendo una continuità locale nelle rappresentazioni.

Dal punto di vista del probing causale, l'effetto di $\mathcal{L}_{\text{ctx}}$ sulle rappresentazioni dell'encoder è sostanziale: le mappe di feature di V-JEPA 2 senza context loss sono «rumorose» e mostrano soltanto struttura spaziale frammentata, mentre aggiungendo $\mathcal{L}_{\text{ctx}}$ emergono mappe coerenti in cui parti semanticamente simili (la testa di un cane, le ruote di un'auto) sono mappate agli stessi componenti PCA. Questo significa che la dipendenza causale tra la posizione spaziale del patch nell'immagine (la variabile t nel nostro framework) e la rappresentazione $E(\text{patch})$ aumenta notevolmente con la context loss.

Quantitativamente, il guadagno in segmentazione semantica lineare — un proxy diretto della forza del probing spaziale — è drammatico: da 22.2 mIoU a 33.9 mIoU su ADE20K per la sola aggiunta di $\mathcal{L}_{\text{ctx}}$, e ulteriori miglioramenti fino a 47.9 mIoU con la pipeline completa (Deep Self-Supervision, VisionMix, Multi-Modal Tokenizer, scaling a ViT-G, cooldown ad alta risoluzione). La stima di profondità lineare su NYUv2 passa da un RMSE di 0.682 (V-JEPA 2) a 0.307 (V-JEPA 2.1 ViT-G), migliorando anche rispetto a DINOv3 con 7 miliardi di parametri (RMSE 0.309).

In termini di probing causale, questo profilo di risultati suggerisce che V-JEPA 2.1 ha sviluppato una forma di *dipendenza causale spaziale* nelle sue rappresentazioni: la posizione 2D del patch causa variazioni sistematiche nella rappresentazione dell'encoder, in modo tale che un probing classifier lineare possa recuperare la geometria 3D della scena direttamente dalle feature congelate. Questa è esattamente la proprietà che la context loss è progettata a indurre: obbligare ogni token contestuale a mantenere informazione sulla propria posizione spazio-temporale relativa ai patch mascherati impedisce che i token contestuali collassino in rappresentazioni globali amorphe — il fenomeno che Darcet et al. (2023) hanno identificato come effetto dei «register tokens» in DINOv2.

12.6 Deep Self-Supervision come regolarizzazione causale multilivello

La seconda innovazione architetturale di V-JEPA 2.1 è la *Deep Self-Supervision*: l'applicazione dell'obiettivo di apprendimento autosupervisionato non soltanto all'output dell'encoder, ma anche a più livelli intermedi. In termini tecnici, si concatenano lungo l'asse dei canali gli output di quattro blocchi intermedi dell'encoder (per ViT-G ai blocchi 12, 24, 36 e 48 su 48 totali), si applica un MLP per ridurne la dimensionalità, e si calcola la perdita sia sui token mascherati sia sui token contestuali a questi quattro livelli.

L'effetto sulla trasmissione dell'informazione causale attraverso la profondità della rete è dimostrabile empiricamente: senza Deep Self-Supervision, esiste un ampio divario tra la performance di probing sull'ultimo strato e quella su quattro strati intermedi (per esempio, 34.9 mIoU vs 39.1 mIoU su ADE20K); con Deep Self-Supervision, il divario si riduce drasticamente (42.0 vs 43.9) e le performance sull'ultimo strato si avvicinano a quelle multilivello. Questo

significa che la Deep Self-Supervision forza l'informazione causale spaziale a propagarsi fino agli strati più profondi della rete, rendendo disponibile per un probing classifier lineare sull'ultimo strato una qualità di informazione che prima era accessibile soltanto aggregando strati intermedi.

Dal punto di vista del probing causale, la Deep Self-Supervision è equivalente a un vincolo di *causality propagation*: imponendo che le strutture causali spaziali siano decodificabili non soltanto all'input e all'output dell'encoder, ma anche agli strati intermedi, si garantisce che le trasformazioni computate dalla rete non «distruggano» informazione causalmente rilevante lungo il percorso computazionale. Questa proprietà è preziosa per l'interpretabilità: un probing causale applicato a strati intermedi può tracciare come l'informazione causale si evolve attraverso la profondità del modello, identificando i livelli in cui avvengono le trasformazioni causalmente rilevanti (per esempio, dove la posizione spaziale raw si trasforma in semantica oggettuale).

12.7 Limiti del probing causale e il problema della spurious correlation

Il probing causale — anche nella versione più rigorosa basata su interventi controllati — non è esente da limitazioni strutturali. La più importante riguarda il fatto che un probing classifier lineare misura l'informazione linearmente decodificabile, non l'informazione causalmente rilevante nel senso più ampio. È perfettamente possibile che uno spazio latente codifichi informazione causalmente determinante in modo non lineare — attraverso interazioni tra dimensioni, curvatura nello spazio latente, o strutture distribuite — che un classificatore lineare non riesce a recuperare. In questo caso, un basso R^2 di probing non implica assenza di dipendenza causale.

Il problema opposto è altrettanto serio: un alto R^2 di probing non garantisce dipendenza causale genuina se lo spazio latente è sufficientemente ricco da codificare correlazioni spurie con alta fedeltà. Il fenomeno delle *spurious correlations* — correlazioni del dataset di training che non riflettono relazioni causali stabili — è documentato ampiamente nella letteratura dell'OOD generalization (Arjovsky et al. 2019, Beery et al. 2018). Un sistema JEPa addestrato su un dataset in cui la rotazione è correlata con il colore dell'oggetto potrebbe sviluppare uno spazio latente dove l'informazione sul colore è decodificabile attraverso lo stesso probing classifier addestrato per decodificare la rotazione, producendo un alto R^2 di equivarianza che riflette una correlazione spurie piuttosto che una dipendenza causale genuina.

Il modo più diretto per separare causalità genuina da correlazione spurea è il test di *causal robustness*: si addestra il probing classifier su un dataset con una struttura di correlazione, e si valuta su un dataset con una struttura di correlazione modificata (dove la correlazione spuria è assente o invertita). Se l' R^2 si degrada significativamente, il classificatore stava sfruttando la

correlazione spuria. Se rimane stabile, la dipendenza codificata è causalmente robusta.

Il benchmark 3DIEBench-OOD costruito nell'ambito di seq-JEPA — con rotazioni campionate dall'intervallo $(-\pi, -\pi/2) \cup (\pi/2, \pi)$, disgiunto dall'intervallo di training $(-\pi/2, \pi/2)$ — è esattamente questo test di robustness causale. I risultati mostrano che tutti i metodi, incluso seq-JEPA, falliscono il test di equivarianza OOD: gli R^2 scendono a valori negativi vicini a -0.7 . Tuttavia, seq-JEPA mostra una degradazione più graduale dell'accuratezza di classificazione rispetto a tutti i baseline: 65.03% contro 63.86% per SimCLR, 55.68% per BYOL, 60.03% per SEN. Questo suggerisce che il meccanismo di aggregazione di seq-JEPA produce una forma di invarianza causale robustezza all'OOD — anche quando l'equivarianza dell'encoder si degrada sotto interventi fuori distribuzione, l'aggregazione multi-vista filtra la variabilità trasformazionale e recupera il contenuto semantico invariante.

Capitolo 13 — TCAV applicato alle rappresentazioni JEPA

13.1 Il framework TCAV: Concept Activation Vectors nello spazio latente

Il Testing with Concept Activation Vectors (TCAV), introdotto da Kim et al. (2018), propone un approccio all'interpretabilità dei modelli profondi che si distingue dal probing classico per la direzione dell'indagine: invece di misurare quali proprietà sono codificate nelle rappresentazioni, TCAV misura *quanto* le previsioni del modello dipendono da concetti definiti esternamente. La differenza è sottile ma fondamentale: il probing chiede «questo concetto c'è?», TCAV chiede «questo concetto conta?».

Formalmente, sia $m : \mathcal{X} \rightarrow \mathbb{R}$ una funzione che restituisce il logit di interesse del modello (per esempio, il logit per la classe «gatto» in un classificatore), sia $E_l : \mathcal{X} \rightarrow \mathbb{R}^d$ il mapping al livello l dell'encoder. Sia C un *concetto* definito da un set di esempi positivi $D_C \subset \mathcal{X}$ (immagini che esemplificano il concetto) e un set di esempi negativi $D_{C'} \subset \mathcal{X}$ (immagini che non esemplificano il concetto). Il *Concept Activation Vector* (CAV) $v_C \in \mathbb{R}^d$ è il vettore normale all'iperpiano di separazione lineare tra le rappresentazioni allo strato l degli esempi positivi e di quelli negativi:

$$v_C = \operatorname{argmin}_{\{v : \|v\|=1\}} \sum_{\{x \in D_C\}} \max(0, 1 - v \cdot E_l(x)) + \sum_{\{x \in D_{C'}\}} \max(0, 1 + v \cdot E_l(x))$$

ossia la normale del classificatore SVM lineare addestrato su $\{E_l(x)\}_{x \in D_C \cup D_{C'}}$ con etichette +1 per D_C e -1 per $D_{C'}$. Una volta calcolato v_C , la sensibilità al concetto C per una data classe k e un dato input x è:

$$\text{TCAV}_{\{C,k\}}(x) = \partial m_k(x) / \partial E_l(x) \cdot v_C$$

dove $m_k(x)$ è il logit del modello per la classe k valutato su x , e il punto è il prodotto scalare tra il gradiente dell'output rispetto alle feature dello strato l e il CAV. Questa quantità misura la direzione di variazione dell'output m_k quando le feature di E_l vengono perturbate nella direzione del concetto C : se il valore è positivo, l'attivazione del concetto C aumenta il logit della classe k ; se negativo, lo diminuisce.

La *TCAV score* aggregata su un dataset di input di classe k è:

$$\text{TCAV_score}(C, k) = |\{x \in D_k : \text{TCAV}_{\{C,k\}}(x) > 0\}| / |D_k|$$

ossia la proporzione di input della classe k per cui il concetto C influenza positivamente la previsione. Se questa proporzione è significativamente diversa da 0.5 (il livello atteso per un concetto casuale irrilevante), il concetto C è genuinamente rilevante per la classe k .

13.3.1 Derivazione del TCAV gerarchico: la formula del coseno tra CAV di livelli diversi e la metrica di coerenza concettuale

Il profilo di CAV descritto nella sezione 13.3 — quattro vettori $v_C^{\{l_1\}}$, $v_C^{\{l_2\}}$, $v_C^{\{l_3\}}$, $v_C^{\{l_4\}}$ calcolati alle profondità $l_1 < l_2 < l_3 < l_4$ dell'encoder — costituisce una sequenza di oggetti nello stesso spazio $\mathbb{R}^d$ (assumendo che la dimensionalità degli strati sia costante, come nei ViT dove ogni strato produce token di dimensione d), ma il suo significato geometrico richiede una cautela fondamentale: due vettori in $\mathbb{R}^d$ non sono confrontabili direttamente se non attraverso una metrica che tenga conto della struttura locale dello spazio latente a ciascuna profondità. La metrica naturale per TCAV è il coseno, perché l'SVM lineare che produce il CAV opera su normalizzazioni delle rappresentazioni, e la direzione del CAV — non la sua magnitudine — è l'oggetto semanticamente rilevante.

Siano $E_{\{l\}}: \mathcal{X} \rightarrow \mathbb{R}^d$ l'embedding prodotto dall'encoder al livello l e $v_C^{\{l\}} = \argmin_{\|v\|=1} L_{\text{SVM}}(v, \{E_l(x)\}_{x \in D_C \cup D_{C'}})$ il CAV del concetto C a quel livello, con la norma a uno garantita dall'addestramento SVM sulla sfera unitaria. La *coerenza gerarchica* del concetto C tra il livello l e il livello $l' > l$ è definita come:

$$\kappa(C, l, l') = v_C^{\{l\}} \cdot v_C^{\{l'\}} / (\|v_C^{\{l\}}\| \cdot \|v_C^{\{l'\}}\|) = v_C^{\{l\}} \cdot v_C^{\{l'\}}$$

dove l'ultima uguaglianza vale perché i CAV sono normalizzati per costruzione. Questa quantità $\kappa(C, l, l') \in [-1, 1]$ misura l'allineamento angolare tra le direzioni concettuali a due profondità: se $\kappa \approx 1$, il concetto è codificato nella stessa direzione a entrambi i livelli; se $\kappa \approx 0$, le direzioni sono ortogonali e il concetto è indipendentemente codificato nei due spazi; se $\kappa \approx -1$, la direzione si inverte — eventualità che indicherebbe una non-monotonicità nell'elaborazione del concetto.

Questa formula, nella sua semplicità, necessita di una derivazione che giustifichi perché il coseno tra vettori di strati diversi abbia significato semantico e non sia un artefatto della metrica euclidea $\mathbb{R}^d$. L'argomento procede per il tramite della struttura di trasferimento tra strati. In un ViT, la propagazione forward dal livello l al livello l' è data da:

$$E_{\{l'\}}(x) = T_{\{l \rightarrow l'\}}(E_l(x))$$

dove $T_{\{l \rightarrow l'\}}$ è la composizione degli $L' - L$ blocchi di self-attention e MLP che separano i due strati. Questa mappa è in generale non lineare e non invertibile. Tuttavia, nel regime lineare locale (linearizzazione di T intorno a una rappresentazione di riferimento $E_l(x_0)$), si può scrivere:

$$E_{\{l'\}}(x) \approx E_{\{l'\}}(x_0) + J_{\{l \rightarrow l'\}}(x_0) \cdot (E_l(x) - E_l(x_0))$$

dove $J_{\{l \rightarrow l'\}}(x_0) = \partial E_{\{l'\}}(x) / \partial E_l(x) |_{\{x=x_0\}} \in \mathbb{R}^{d \times d}$ è lo Jacobiano della mappa di propagazione valutato alla rappresentazione di riferimento. Il CAV al livello l' è determinato dalla geometria delle rappresentazioni $\{E_{\{l'\}}(x)\}_{x \in D_C \cup D_{C'}}$. Usando la linearizzazione:

$$v_C^{\{l'\}} \approx (J_{\{l \rightarrow l'\}}^T)^{-1} \cdot v_C^{\{l\}} / \| (J_{\{l \rightarrow l'\}}^T)^{-1} \cdot v_C^{\{l\}} \|$$

ovvero, se la mappa di propagazione $T_{\{l \rightarrow l'\}}$ è localmente isometrica (conserva le distanze euclidee), allora $J_{\{l \rightarrow l'\}}$ è ortogonale ($J^T J = I$) e si ha:

$$v_C^{\{l'\}} \approx J_{\{l \rightarrow l'\}} \cdot v_C^{\{l\}} / \| J_{\{l \rightarrow l'\}} \cdot v_C^{\{l\}} \|^2$$

In questo caso, la coerenza gerarchica diventa:

$$\kappa(C, l, l') = v_C^{\{l\}} \cdot v_C^{\{l'\}} = v_C^{\{l\}} \cdot (J_{\{l \rightarrow l'\}} \cdot v_C^{\{l\}}) / \| J_{\{l \rightarrow l'\}} \cdot v_C^{\{l\}} \|^2$$

che può essere interpretata come il coseno dell'angolo di rotazione subito dal CAV durante la propagazione attraverso i blocchi dell'encoder. Per una mappa localmente isometrica, questo angolo misura la *rotazione concettuale* indotta dall'elaborazione gerarchica: il concetto ruota nello spazio d -dimensionale man mano che l'encoder integra l'informazione locale in strutture di ordine superiore.

Nel caso più generale (non isometrico), il Jacobiano $J_{\{l \rightarrow l'\}}$ ha valori singolari $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d$ che esprimono la deformazione dello spazio latente. Il CAV

$v_C^{\wedge\{l\}}$ si decompone sulla base delle direzioni singolari destre: $v_C^{\wedge\{l\}} = \sum_i \alpha_i \hat{e}_i$, dove $\{\hat{e}_i\}$ sono i vettori singolari destri di $J\{l \rightarrow l'\}$. La mappa produce $J_{-}\{l \rightarrow l'\} \cdot v_C^{\wedge\{l\}} = \sum_i \alpha_i \sigma_i f_i$, con $\{f_i\}$ i vettori singolari sinistri. La coerenza gerarchica del CAV è:

$$\kappa(C, l, l') \approx [v_C^{\wedge\{l\}} \cdot J_{-}\{l \rightarrow l'\} \cdot v_C^{\wedge\{l\}}] / \|J_{-}\{l \rightarrow l'\} \cdot v_C^{\wedge\{l\}}\| = [\sum_i \alpha_i^2 \sigma_i] / \|\sum_i \alpha_i \sigma_i f_i\|$$

Questa espressione rivela una proprietà strutturale importante: la coerenza è elevata quando il CAV è allineato con le direzioni singolari dominanti di J — cioè quando la direzione concettuale è nelle sottospazio che viene amplificato dalla mappa di propagazione. In altri termini, un concetto che è «proiettato in avanti» dall'encoder mantiene una direzione coerente attraverso i livelli, mentre un concetto che si trova nello spazio nullo o quasi-nullo di J si dissolve nella propagazione.

Per il TCAV gerarchico in V-JEPA 2.1, questa derivazione ha due implicazioni operative verificabili. La prima riguarda la previsione per i concetti di basso livello come i bordi: ci si aspetta che $\kappa(C_{\text{bordo}}, l_1, l_2)$ sia elevato (le direzioni dei bordi sono stabili tra blocco 12 e blocco 24) ma che $\kappa(C_{\text{bordo}}, l_1, l_4)$ sia basso (all'uscita del blocco 48, la codifica dei bordi primitivi ha subito un processo di integrazione che ne cambia la direzione di rappresentazione). La seconda implicazione riguarda i concetti semantici di alto livello come la categoria di oggetto: ci si aspetta il pattern opposto, con $\kappa(C_{\text{sem}}, l_1, l_3)$ basso e $\kappa(C_{\text{sem}}, l_3, l_4)$ elevato — il concetto semantico emerge e si stabilizza negli strati finali.

La matrice di coerenza gerarchica completa $K(C) \in \mathbb{R}^{4 \times 4}$ con $K(C)_{ij} = \kappa(C, l_i, l_j)$ costituisce dunque un *fingerprint geometrico* del concetto C nell'encoder: una struttura a bassa coerenza nelle celle fuori-diagonale superiori (l_3 vs l_1 , l_1 vs l_4) indica un concetto di basso livello che si integra rapidamente; una struttura a elevata coerenza nelle celle fuori-diagonale inferiori (l_3 vs l_4 , l_2 vs l_4) indica un concetto semantico che emerge progressivamente. L'effetto della Deep Self-Supervision di V-JEPA 2.1 sulla matrice K è prevedibile: forzando ogni livello intermedio a mantenere informazione sufficiente per il proprio obiettivo di predizione, la Deep Self-Supervision dovrebbe aumentare i valori di $\kappa(C, l_i, l_j)$ per concetti spaziali a media scala (relazioni tra regioni, struttura di sfondo/primo piano), che sono il target esplicito della perdita multi-livello, mentre dovrebbe lasciare invariata o ridurre la coerenza per i concetti di livello più fine (bordi primitivi) che non sono il focus dell'obiettivo ausiliario.

Si noti infine che la formula del coseno tra CAV di livelli diversi è verificabile sperimentalmente senza richiedere accesso alle rappresentazioni interne del modello a runtime: è sufficiente fare probing alle quattro profondità sull'encoder frozen, calcolare i CAV per un insieme di concetti di interesse, e misurare le matrici $K(C)$. Questo la rende uno strumento di analisi dell'interpretabilità accessibile con la stessa infrastruttura sperimentale richiesta dal probing classico.

13.2 Il calcolo dei CAV negli spazi JEPA biforcati

Gli spazi latenti JEPA presentano una caratteristica strutturale inusuale rispetto agli spazi latenti su cui TCAV è stato originariamente sviluppato: la biforcazione architetturale tra rappresentazioni dell'encoder z_i (equivarianti alle trasformazioni) e rappresentazione aggregata z_{AGG} (invariante alle trasformazioni). Questa biforcazione cambia l'interpretazione dei CAV in modo non triviale.

Per le rappresentazioni dell'encoder $z_i = E(x_i)$, i CAV calcolati su concetti visivi locali — la tessitura di una superficie, l'orientamento di un bordo, la posizione di un keypoint — hanno un significato geometrico preciso nello spazio latente: essi identificano le direzioni lungo cui varia la risposta a caratteristiche locali dell'immagine. Poiché E è addestrato a essere equivariante alle trasformazioni, le rappresentazioni z_i cambiano in modo sistematico al variare della trasformazione applicata all'input. I CAV calcolati su esempi della stessa categoria con trasformazioni diverse catturano la *struttura di equivarianza* dello spazio latente: essi descrivono la geometria delle orbite delle trasformazioni nello spazio latente.

Per la rappresentazione aggregata z_{AGG} , i CAV hanno un significato diverso: le rappresentazioni z_{AGG} sono costruttivamente invarianti alle trasformazioni per come l'aggregatore funge da integrazione multi-vista, e i CAV calcolati su z_{AGG} riflettono quindi strutture concettuali di ordine superiore che sono stabili attraverso le variazioni trasformazionali. Un CAV per il concetto «rettangolarità» calcolato sullo spazio z_{AGG} non è influenzato dalla rotazione dell'oggetto — la proprietà «rettangolare» è astratta rispetto all'orientamento specifico — mentre un CAV calcolato su z_i riflette la rettangolarità percepita dal punto di vista specifico da cui è stata catturata la vista i .

La distinzione ha conseguenze pratiche per il calcolo. Siano $D_C^{\{rot=0^\circ\}}$ gli esempi del concetto C fotografati da una prospettiva frontale e $D_C^{\{rot=90^\circ\}}$ gli stessi oggetti fotografati con una rotazione di 90° . Nello spazio z_i , i due CAV calcolati su questi due dataset saranno generalmente diversi — perché E è equivariante e mappa le due prospettive in posizioni diverse dello spazio latente. Nello spazio z_{AGG} (computato aggregando entrambe le viste), i due CAV saranno invece simili o identici — perché l'aggregazione ha rimosso la dipendenza dalla prospettiva specifica.

Formalmente, sia $v_C^{\{rot=0^\circ\}}$ il CAV del concetto C calcolato sullo spazio z_i a partire da esempi con rotazione 0° , e $v_C^{\{rot=90^\circ\}}$ il corrispondente CAV a partire da esempi con rotazione 90° . L'*equivarianza del CAV* è misurata da:

$$\text{sim_CAV}(C, E) = |v_C^{\{rot=0^\circ\}} \cdot v_C^{\{rot=90^\circ\}}| / (\|v_C^{\{rot=0^\circ\}}\| \cdot \|v_C^{\{rot=90^\circ\}}\|)$$

Se $\text{sim_CAV}(C, E) \approx 0$, le due prospettive producono direzioni concettuali quasi ortogonali nello spazio E : il concetto è codificato in modo vista-dipendente. Se $\text{sim_CAV}(C, E) \approx 1$, le due prospettive producono la stessa direzione concettuale: il concetto è codificato in modo vista-indipendente. Sperimentalmente, per un encoder JEPA equivariante come quello di seq-JEPA si attende $\text{sim_CAV}(C, E) \approx 0$ sullo spazio z_i (alta equivarianza) e $\text{sim_CAV}(C, E) \approx 1$ sullo spazio z_{AGG} (alta invarianza). Questa predizione è verificabile direttamente attraverso il calcolo dei CAV sulle rappresentazioni di seq-JEPA.

13.3 TCAV gerarchico e Deep Self-Supervision in V-JEPA 2.1

L'architettura di Deep Self-Supervision in V-JEPA 2.1, che applica l'obiettivo di apprendimento a quattro strati intermedi dell'encoder, offre un contesto naturale per il calcolo gerarchico dei CAV. Invece di calcolare un singolo CAV per l'intero modello, si può calcolare un profilo di CAV per il medesimo concetto a quattro profondità diverse: l'uscita del blocco 12, del blocco 24, del blocco 36 e del blocco 48 (l'output finale). La traiettoria di questo profilo — come il CAV del concetto C si trasforma attraverso la profondità del modello — rivela la struttura computazionale del processo di comprensione del concetto.

La previsione teorica, derivabile dai principi dell'apprendimento gerarchico delle feature, è che i concetti di basso livello (bordi, texture, gradiente di luminanza) siano codificati negli strati bassi e che la loro direzione CAV decresca di magnitudine negli strati alti man mano che queste informazioni vengono integrate in strutture di ordine superiore. I concetti di alto livello (semantica categoriale, relazioni spaziali tra oggetti, proprietà fisiche come la rigidità) dovrebbero avere invece CAV di magnitudine crescente con la profondità.

Senza Deep Self-Supervision, le rappresentazioni degli strati intermedi di V-JEPA 2 non sono direttamente supervisionate dall'obiettivo JEPA: la perdita agisce soltanto sull'output finale. In questo caso, ci si attende che i CAV degli strati intermedi siano meno stabili e meno informativi rispetto ai CAV dell'output finale. Con Deep Self-Supervision, ogni strato intermedio è costretto a mantenere informazione spazialmente strutturata sufficiente per la perdita multi-livello, il che dovrebbe produrre CAV più stabili e interpretabili a ogni profondità.

Questa previsione è coerente con il risultato empirico di Mur-Labadia et al. riportato nella Tabella 12 del paper di V-JEPA 2.1: senza Deep Self-Supervision, l'uso di quattro strati intermedi per il probing di segmentazione (4-Layers: 39.1 mIoU su ADE20K) produce un guadagno significativo rispetto all'uso del solo strato finale (Last-Layer: 34.9 mIoU). Con Deep Self-Supervision, il divario si riduce a 43.9 mIoU (4-Layers) vs 42.0 mIoU (Last-Layer): l'informazione rilevante si è propagata verso lo strato finale, rendendo meno necessaria l'aggregazione multilivello per il probing.

13.4 Concetti come vincoli causali nello spazio latente JEPA

La connessione tra TCAV e il framework del probing causale del Capitolo 12 diventa produttiva quando si interpretano i CAV come vettori che *spannano* le direzioni causalmente rilevanti nello spazio latente. Un CAV v_C calcolato su esempi del concetto C definisce una direzione nello spazio latente lungo cui la variazione produce un effetto sistematico sull'output del modello. Se questo effetto è sistematico e robusto — se la TCAV score è significativamente diversa da 0.5 su dataset con diverse strutture di correlazione — allora v_C è un vettore *causalmente rilevante* nel senso del Capitolo 12.

Il valore aggiunto di TCAV rispetto al probing classico è che esso permette di testare concetti definiti dall'utente, non soltanto variabili di fattore predefinite dal processo generativo del dataset. Nell'ambito dell'analisi di un sistema JEPA addestrato su video egocentrici per la robotica, si potrebbe definire un concetto come «superficie su cui posare un oggetto» attraverso un set di immagini positive (superfici piane, tavoli, pavimenti) e negative (superfici irregolari, aria, sfondo). Il CAV di questo concetto nello spazio z_{AGG} di V-JEPA 2.1 rivelerebbe se le feature dell'encoder catturano questa proprietà *affordance* — e la TCAV score su una classe di azioni come «posare» direbbe se il modello sta effettivamente usando questa informazione per le sue previsioni.

Questa modalità di analisi è particolarmente rilevante per il programma di allineamento del sistema JEPA a valori e obiettivi umani: permette di verificare non soltanto che il sistema codifichi certi concetti (probing), ma che li usi nelle proprie decisioni (TCAV). La distinzione tra codifica e utilizzo è critica: un sistema potrebbe codificare informazione sull'intenzione dell'utente nei propri stati latenti senza usarla nelle proprie previsioni, o potrebbe usare proxy superficiali (caratteristiche correlate con l'intenzione ma non causalmente rilevanti) al posto dell'intenzione genuina.

13.5 La stabilità dei CAV e il problema della robustezza all'intervento

Un problema aperto nel calcolo dei CAV è la loro stabilità rispetto a perturbazioni del dataset di definizione: se si sostituisce il 10% degli esempi positivi del concetto C con esempi diversi, il CAV cambia significativamente? Kim et al. (2018) propongono un test di significatività statistica basato sul confronto della TCAV score su esempi positivi del concetto contro la TCAV score su concetti randomizzati (CAV calcolati su set di esempi casuali). Se la TCAV score del concetto reale è significativamente maggiore di quella dei concetti casuali, il concetto è interpretato come genuinamente rilevante.

Nel contesto degli spazi JEPA biforcati, la stabilità dei CAV ha implicazioni diverse per z_i e z_{AGG} . Nello spazio z_i , i CAV sono instabili rispetto a perturbazioni della prospettiva di vista degli esempi — perché z_i è equivariante, la sostituzione di esempi frontali con esempi ruotati cambia la

direzione del CAV in modo sistematico. Nello spazio z_AGG , i CAV dovrebbero essere molto più stabili, perché l'aggregazione multi-vista ha rimosso la dipendenza dalla prospettiva specifica: sostituire esempi frontali con esempi ruotati negli stessi oggetti produce z_AGG simili, e quindi CAV simili.

Questa previsione fornisce un test operativo per verificare la struttura di invarianza di un encoder JEPA: si calcolano i CAV per lo stesso concetto con diverse sottocomposizioni degli esempi positivi (diverse prospettive, diverse condizioni di illuminazione, diverse occlusioni parziali), si misura la deviazione standard dei coseni tra i CAV risultanti, e si confronta questa deviazione tra spazio z_i e spazio z_AGG . Un encoder con alta invarianza dovrebbe mostrare deviazione standard bassa su z_AGG e alta su z_i . Questo test è complementare all' R^2 di probing causale discusso nel Capitolo 12: quello misura la forza della dipendenza, questo ne misura la robustezza alle variazioni non-causali.

13.6 Equivarianza come struttura dei CAV nello spazio latente

Nell'ambito del framework seq-JEPA, i risultati di probing equivariante offrono un modo diretto per calcolare i CAV delle trasformazioni. Una trasformazione t — per esempio una rotazione $SO(3)$ di angolo θ attorno a un asse $\hat{n}$ — definisce naturalmente una famiglia di concetti parametrica: il concetto $C_\theta = \{(x, t(x)) : x \in \mathcal{X}\}$ è l'insieme delle coppie (vista originale, vista ruotata di θ). Il CAV di questo concetto calcolato nello spazio z_i è il vettore $v_{\{C_\theta\}}$ che separa le rappresentazioni delle viste originali da quelle ruotate di θ .

Se l'encoder è *perfettamente equivariante* a $SO(3)$ nel senso della definizione 2 del paper di Ghaemi et al. — cioè $\exists u_a$ tale che $f(x_2) = u_a(f(x_1))$ dove a è l'azione che porta da x_1 a x_2 — allora la struttura del gruppo di rotazioni $SO(3)$ dovrebbe essere riflessa nella struttura dei CAV nello spazio latente. In particolare, per la composizione di rotazioni $SO(3)$ vale la regola:

$$v_{\{C_{\{\theta_1 \circ \theta_2\}}\}} \approx R_u \cdot (v_{\{C_{\{\theta_1\}}\}} + v_{\{C_{\{\theta_2\}}\}})$$

dove R_u è la rappresentazione lineare in Z del gruppo di rotazioni, e la somma dei CAV approssima il CAV della rotazione composta. Questa proprietà — che i CAV si comportino come vettori di uno spazio di rappresentazione lineare del gruppo di simmetria — è la condizione che renderebbe lo spazio latente di un encoder JEPA genuinamente equivariante nel senso algebrico più stringente, non soltanto nel senso statistico misurato dall' R^2 .

I risultati di seq-JEPA con ablazione dell'action conditioning (Table 4 di Ghaemi et al.) mostrano che senza conditioning l' R^2 di equivarianza scende da 0.71 a 0.29, pur mantenendo l'accuratezza di classificazione (87.36% vs 87.41%). Questo significa che la struttura del gruppo di trasformazioni non è riflessa nello spazio latente dell'encoder quando il conditioning sulle azioni viene rimosso: i CAV delle trasformazioni potrebbero esistere (l'encoder codifica ancora alcune informazioni sulla rotazione), ma non obbediscono alla

struttura algebrica del gruppo. Il conditioning sulle azioni durante il training è il meccanismo che impone questa struttura algebrica sui CAV — obbligando l'encoder a mantenere informazione sufficiente sulla trasformazione perché il predittore possa predire la vista successiva condizionato sull'azione, lo si costringe a organizzare le proprie feature nella geometria del gruppo di trasformazione.

13.6.1 Prova formale della stabilità dei CAV rispetto agli interventi sul DAG JEPA

La sezione 13.5 ha descritto qualitativamente la stabilità dei CAV come funzione della struttura di invarianza dell'encoder — distinguendo il comportamento di z_i (equivariante e quindi instabile rispetto alla prospettiva) da quello di z_{AGG} (invariante e quindi stabile). La sezione 13.6 ha connesso la struttura dei CAV all'equivarianza del gruppo di simmetria attraverso la formula di composizione dei CAV. Il presente paragrafo completa questo quadro dimostrando formalmente la stabilità dei CAV rispetto a una classe più ampia di perturbazioni: gli interventi causali sul DAG latente del modello JEPA, nel senso del do-calcolo di Pearl.

Definizioni e notazione. Sia $Z = (Z_1, Z_2, \dots, Z_m)$ il vettore di variabili latenti del DAG generativo JEPA, con struttura causale descritta dal grafo $G = (V, E)$ e dalle equazioni strutturali $Z_i = f_i(Z_{\text{pa}(i)}, N_i)$, dove $\text{pa}(i)$ sono i genitori di Z_i in G e N_i è il rumore esogeno. Un concetto C è definito come una partizione binaria dello spazio osservativo $\mathcal{X}$ tramite i dataset D_C e $D_{C'}$. Il CAV $v_C^{\{l\}}$ al livello 1 è il vettore del classificatore SVM addestrato sulle rappresentazioni $\{E_l(x)\}_{x \in D_C \cup D_{C'}}$.

Un *intervento causale* $\text{do}(Z_i = \tilde{z}_i)$ nel senso di Pearl modifica il DAG G producendo il grafo mutilato $G_{\{X_i\}}$ in cui tutti gli archi entranti in Z_i vengono rimossi e Z_i viene fissato al valore $\tilde{z}_i$. Nella generazione dei dati di definizione del concetto, un intervento $\text{do}(Z_i = \tilde{z}_i)$ corrisponde operativamente a condizionare la selezione degli esempi positivi D_C a campioni generati con il fattore latente Z_i fissato a $\tilde{z}_i$.

Teorema (Stabilità dei CAV sotto intervento). Sotto le seguenti assunzioni: (A1) Il DAG generativo G è causalmente corretto: il processo generativo degli esempi D_C e $D_{C'}$ è consistente con le equazioni strutturali. (A2) Il concetto C è d-separato da Z_i in G dato l'insieme di osservazioni $\mathcal{X}$: $Z_C \perp_G Z_i \mid \mathcal{X}$, dove Z_C indica l'insieme di variabili latenti che causano le differenze tra D_C e $D_{C'}$. (A3) L'encoder E_l è fedele alla struttura di G nel senso che le dipendenze statistiche tra le rappresentazioni $E_l(x)$ riflettono le dipendenze causali tra le variabili latenti che generano x .

Allora il CAV del concetto C è stabile rispetto all'intervento $\text{do}(Z_i = \tilde{z}_i)$:

$$v_C^{\{l\}} \mid \{\text{do}(Z_i = \tilde{z}_i)\} \approx v_C^{\{l\}} \mid \{\text{obs}\}$$

nel senso che la distanza coseno tra i due CAV soddisfa:

$$1 - \kappa(v_{C^{\wedge}\{I\}} | \{do(Z_i = \tilde{z}_i)\}, v_{C^{\wedge}\{I\}} | \{obs\}) \leq O(\|Cov(E_l(X), Z_i | Z_C)\| / \|\text{Var}(E_l(X) | Z_C)\|)$$

ossia la variazione del CAV è controllata dal rapporto tra la covarianza residua tra la rappresentazione $E_l(X)$ e la variabile intervenuta Z_i (condizionata sui fattori causali del concetto Z_C) e la varianza delle rappresentazioni spiegata da Z_C .

Dimostrazione. La dimostrazione procede per quattro passi.

Passo 1: rappresentazione della distribuzione interventale sulle rappresentazioni. Sotto l'assunzione (A1), la distribuzione osservazionale delle rappresentazioni $E_l(x)$ per $x \in D_C$ è:

$$p(E_l(x) | x \in D_C) = \int p(E_l(x) | Z = z) \cdot p(Z = z | x \in D_C) dz$$

Sotto l'intervento $do(Z_i = \tilde{z}_i)$, la distribuzione diventa:

$$p(E_l(x) | x \in D_C, do(Z_i = \tilde{z}_i)) = \int p(E_l(x) | Z_{-i} = z_{-i}, Z_i = \tilde{z}_i) \cdot p(Z_{-i} = z_{-i} | x \in D_C, Z_i = \tilde{z}_i) dz_{-i}$$

dove Z_{-i} indica tutte le variabili latenti tranne Z_i .

Passo 2: uso della d-separazione. Per l'assunzione (A2), $Z_C \perp_G Z_i | \mathcal{D}$. Questo implica, per il criterio di Markov globale sui DAG, che la distribuzione dei fattori causali del concetto Z_C è invariante all'intervento su Z_i dato l'insieme di osservazioni compatibili con D_C :

$$p(Z_C | x \in D_C, do(Z_i = \tilde{z}_i)) = p(Z_C | x \in D_C)$$

ovvero l'intervento su Z_i non modifica la distribuzione dei fattori che definiscono il concetto C . Questo è il nucleo dell'argomentazione: la d-separazione garantisce che il «segnale» causale del concetto sia ortogonale, nel grafo, alla variabile intervenuta.

Passo 3: impatto sulla distribuzione delle rappresentazioni. Per l'assunzione (A3), la struttura di dipendenza tra $E_l(x)$ e le variabili latenti è:

$$E_l(x) = \mu_l(Z_C) + \Delta_l(Z_i, Z_{-C,-i}) + \eta_l$$

dove $\mu_l(Z_C)$ è il componente delle rappresentazioni spiegato dai fattori causali del concetto, Δ_l è il componente spiegato dalle variabili latenti non-causali (incluso Z_i), e η_l è rumore residuo. Sotto l'assunzione di d-separazione (Passo 2), la componente $\mu_l(Z_C)$ ha la stessa distribuzione nei due regimi (osservazionale e interventale). La variazione delle rappresentazioni sotto l'intervento proviene esclusivamente da $\Delta_l(Z_i = \tilde{z}_i, \cdot)$ rispetto a $\Delta_l(Z_i, \cdot)$ nella distribuzione osservazionale.

Passo 4: variazione del CAV e bound del coseno. Il CAV $v_C^{\{ \}$ è la normale dell'iperpiano ottimale dell'SVM. Per piccole perturbazioni della distribuzione dei dati di training, la variazione della normale dell'iperpiano SVM è controllata dalla variazione della matrice di covarianza dei dati. In particolare, dalla teoria delle perturbazioni per gli SVM a margine massimo, la variazione del vettore normale sotto una perturbazione $\Delta\Omega$ della distribuzione dei dati è:

$$\delta v_C^{\{ \}} \approx (\Pi K)^{\dagger} \cdot \Delta\Omega \cdot \alpha^*$$

dove ΠK è la matrice del kernel valutata sui vettori di supporto, $(\cdot)^{\dagger}$ è la pseudo-inversa, e α^* sono i moltiplicatori ottimali di Lagrange. Per un SVM lineare (kernel $K(z, z') = z \cdot z'$), questo si riduce a:

$$\delta v_C^{\{ \}} \approx \text{Cov}_{\{\text{osservazionale}\}}(E_l(X), Z_i \mid Z_C) \cdot \gamma_i$$

dove γ_i è un vettore di scala determinato dalla distribuzione interventale di Z_i . La norma della perturbazione del CAV è dunque:

$$\|\delta v_C^{\{ \}}\| \leq \|\text{Cov}(E_l(X), Z_i \mid Z_C)\|_F \cdot \|\gamma_i\|_2$$

La variazione del coseno tra il CAV originale e il CAV perturbato è:

$$1 - \kappa(v_C^{\{ \}} \mid \{do\}, v_C^{\{ \}} \mid \{obs\}) = 1 - \cos(\text{angle}(v_C^{\{ \}} \mid \{do\}, v_C^{\{ \}} \mid \{obs\})) \approx \|\delta v_C^{\{ \}}\|^2 / (2 \|v_C^{\{ \}}\|^2) \leq O(\|\text{Cov}(E_l(X), Z_i \mid Z_C)\|^2 / \|v_C^{\{ \}}\|^2)$$

Poiché $\|v_C^{\{ \}}\|^2 = \|\mu_l(Z_C)\|^2 / \|\text{Var}(E_l(X) \mid Z_C)\| + O(\|\eta_l\|^2)$, si ottiene il bound enunciato nel teorema. $\square$

Interpretazione e conseguenze. Il teorema fornisce una garanzia formale del seguente tipo: il CAV del concetto C è stabile rispetto all'intervento $do(Z_i = \tilde{z}_i)$ nella misura in cui $E_l(x)$ è *causalmente indipendente* da Z_i dato i fattori del concetto Z_C . Quando l'encoder JEPA soddisfa l'assunzione (A3) — fedeltà alla struttura causale del DAG latente — la stabilità del CAV è direttamente controllata dalla struttura causale del modello generativo.

Questa garanzia ha tre corollari rilevanti per la verifica della struttura causale degli encoder JEPA.

Il *primo corollario* stabilisce che se un encoder non è causalmente fedele — ovvero se le sue rappresentazioni confondono i fattori causali del concetto con variabili intervenibili non causali — allora i CAV saranno instabili rispetto agli interventi. Questo fornisce un test empirico della fedeltà causale: si calcolano i CAV con dataset D_C raccolti in condizioni interventali controllate (variando sistematicamente Z_i), si misura la coerenza coseno tra i CAV risultanti, e si verifica se essa è compatibile con il bound del teorema. Un'instabilità elevata indica che l'encoder sta confondendo Z_i con Z_C — ovvero sta codificando correlazioni spurie come se fossero struttura causale.

Il *secondo corollario* riguarda la relazione tra la struttura di equivarianza di seq-JEPA e la stabilità dei CAV. La decomposizione $z_i = E(x_i)$ è progettata per essere equivariante alle trasformazioni del gruppo G_T . Le trasformazioni del gruppo — rotazioni, traslazioni, cambi di prospettiva — corrispondono a interventi su variabili latenti di tipo pose o viewpoint nel DAG generativo. Il teorema afferma che se il concetto C è causalmente separato da questi fattori di posa (il che è garantito dalla struttura di z_{AGG} , che aggrega e marginalizza rispetto alla prospettiva), allora i CAV calcolati su z_{AGG} sono stabili rispetto agli interventi di trasformazione. Questo è il risultato della sezione 13.5 — la previsione di deviazione standard bassa su z_{AGG} e alta su z_i — ottenuto ora come corollario del teorema di stabilità causale piuttosto che come osservazione euristica.

Il *terzo corollario* riguarda l'applicazione al probing delle strutture affordance. Se il concetto «superficie di appoggio» è causalmente separato dalle variabili di illuminazione e di prospettiva nel DAG generativo della scena (cosa plausibile perché l'affordance di una superficie è una proprietà intrinseca dell'oggetto, non della sua presentazione), allora il CAV di questo concetto su z_{AGG} dovrebbe essere stabile rispetto a variazioni di illuminazione e di punto di vista degli esempi di training. Il teorema quantifica questa stabilità in termini della covarianza residua $Cov(E_1(X), Z_{\text{illuminazione}} | Z_{\text{affordance}})$, che è piccola nella misura in cui l'encoder ha appreso a separare l'informazione di illuminazione dall'informazione di affordance. La verifica empirica di questa predizione — misurando la coerenza dei CAV di affordance tra dataset raccolti in condizioni di illuminazione diverse — costituisce un test indiretto della qualità della struttura causale appresa dall'encoder.

Limiti della prova. La dimostrazione si appoggia all'assunzione (A3) di fedeltà causale dell'encoder, che è un'assunzione sullo stato di addestramento del modello e non una proprietà garantita dalla sola architettura JEPA. In pratica, la fedeltà causale è una proprietà emergente dell'ottimizzazione che si avvicina quanto più il processo di apprendimento segue i principi del Capitolo 17 — regolarizzazione della struttura causale attraverso l'ELBO causale e la penalità SIGReg. Il teorema stabilisce dunque una condizione *sufficiente* per la stabilità dei CAV (che l'encoder sia causalmente fedele), non una condizione necessaria: un encoder potrebbe produrre CAV stabili anche senza fedeltà causale completa, purché le componenti non-causali delle sue rappresentazioni siano cancellate dall'SVM durante il calcolo del CAV. Tuttavia, in quel caso la stabilità sarebbe fragile — dipendente dalla struttura specifica del dataset di definizione del concetto e non garantita rispetto a nuove condizioni di intervento — mentre la stabilità garantita dal teorema è robusta rispetto a qualsiasi intervento sul DAG che preservi la d-separazione.

Capitolo 14 — Sparse Autoencoders come strumento di decomposizione degli spazi JEPA

14.1 Il problema della polisemia e la superposizione delle feature

Il probing e TCAV condividono un presupposto implicito che merita di essere reso esplicito: entrambi assumono che i concetti rilevanti per il modello siano allineati con le direzioni dello spazio latente in modo tale che un classificatore o regressore lineare possa recuperarli. Questa assunzione è violata in modo sistematico nei trasformatori profondi — inclusi gli encoder JEPA — a causa del fenomeno della *superposizione delle feature* (Elhage et al. 2022): le reti neurali tendono a rappresentare un numero di feature molto superiore alla dimensionalità dei loro spazi di attivazione, usando combinazioni lineari sovrapposte di direzioni come «codice» per feature distinte. Il risultato è che ciascuna direzione nello spazio latente corrisponde a più feature simultaneamente (polisemia), e ciascuna feature è distribuita su più direzioni (superposizione). Un probing classifier lineare su uno spazio polisemico produce pertanto probing vettori che catturano miscele di feature anziché feature pure, rendendo difficile l'interpretazione.

Gli *Sparse Autoencoders* (SAE) sono stati proposti come soluzione a questo problema: invece di cercare di interpretare le direzioni dello spazio latente dell'encoder originale, si addestra un autoencoder con un vincolo di sparsità sulle attivazioni a ricostruire le rappresentazioni latenti dell'encoder, sperando che gli elementi del dizionario dell'autoencoder corrispondano a feature monosemantiche — feature che rappresentano un unico concetto e vengono attivate soltanto in risposta a input che lo esemplificano.

Formalmente, sia $z \in \mathbb{R}^d$ una rappresentazione latente dell'encoder JEPA. Uno Sparse Autoencoder è una coppia (enc, dec) dove $\text{enc} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ (con $n \gg d$ tipicamente) e $\text{dec} : \mathbb{R}^n \rightarrow \mathbb{R}^d$, parametrizzate da:

$$\text{enc}(z) = \sigma(W_{\text{enc}} \cdot z + b_{\text{enc}}) \quad \text{dec}(\hat{h}) = W_{\text{dec}} \cdot \hat{h} + b_{\text{dec}}$$

dove σ è una funzione di attivazione che favorisce la sparsità (tipicamente ReLU o JumpReLU, la versione con soglia apprendibile introdotta da Rajamanoharan et al. 2024), $\hat{h} = \text{enc}(z) \in \mathbb{R}^n$ è il vettore di attivazioni sparse, e $W_{\text{dec}} \in \mathbb{R}^{d \times n}$ è la matrice dei vettori del dizionario (con colonne normalizzate a norma unitaria). L'obiettivo di training è:

$$\mathcal{L}_{\text{SAE}}(z) = \|z - \text{dec}(\text{enc}(z))\|^2 + \alpha \cdot \|\text{enc}(z)\|_1$$

dove il primo termine è l'errore di ricostruzione e il secondo è il termine di regolarizzazione L_1 che promuove la sparsità delle attivazioni. Il parametro α bilancia il trade-off tra fedeltà della ricostruzione e sparsità delle feature.

14.2 Adattamento degli SAE agli spazi latenti JEPa biforcati

L'applicazione degli SAE agli spazi latenti JEPa richiede di considerare la biforcazione strutturale tra z_i e z_{AGG} . I due spazi hanno caratteristiche statistiche diverse che influenzano il comportamento degli SAE.

Lo spazio dell'encoder z_i è *equivariante* rispetto alle trasformazioni: le rappresentazioni di input trasformati sono sistematicamente diverse dalle rappresentazioni degli input originali, in modo dipendente dalla trasformazione applicata. Questo significa che la distribuzione $p(z_i)$ è tipicamente multimodale e strutturata geometricamente: le viste dello stesso oggetto con diverse rotazioni formano «orbite» nello spazio latente, e la distribuzione all'interno di ciascuna orbita riflette la struttura del gruppo di trasformazione. Un SAE addestrato su $p(z_i)$ dovrebbe sviluppare feature (vettori del dizionario) che corrispondono a proprietà locali dell'input — tessitura, gradiente, orientamento locale — che si attivano in modo sistematico al variare della prospettiva.

Lo spazio aggregato z_{AGG} è *invariante* per costruzione: le rappresentazioni di sequenze di viste dello stesso oggetto con diverse rotazioni convergono verso punti dello spazio latente che riflettono la categoria e le proprietà stabili dell'oggetto, non la prospettiva specifica. Questo significa che la distribuzione $p(z_{AGG})$ è più compatta e monomodale per ciascuna categoria di oggetti, e un SAE addestrato su $p(z_{AGG})$ dovrebbe sviluppare feature (vettori del dizionario) che corrispondono a proprietà semantiche di alto livello — la forma globale dell'oggetto, la sua funzione, la sua categoria.

Per il SAE applicato allo spazio z_i , si propone la seguente modifica architetturale che sfrutta la struttura di equivarianza: si parametrizza il dizionario W_{dec} con colonne organizzate in *orbite* rispetto al gruppo di trasformazione G . Formalmente, per ciascuna feature «radice» $w_0 \in \mathbb{R}^d$, si generano le feature corrispondenti alle sue rotazioni $g(w_0)$ per $g \in G$, e si costruisce il dizionario come unione di tutte le orbite. L'idea è che le feature di una stessa orbita corrispondano alla stessa proprietà visiva locale osservata da diverse prospettive, e che il pattern di attivazione di un'intera orbita codifichi la prospettiva specifica da cui è stata osservata la proprietà.

Questo SAE equivariante può essere addestrato aggiungendo al training objective standard un termine di regolarizzazione che impone la struttura di orbita:

$$\mathcal{L}_{SAE_equiv}(z) = \|z - \text{dec}(\text{enc}(z))\|^2 + \alpha \cdot \|\text{enc}(z)\|_1 + \beta \cdot \sum_k \|\text{enc}_k(z) - \text{enc}_{\{g(k)\}}(g \cdot z)\|^2$$

dove $\text{enc}_k(z)$ è l'attivazione della feature k per l'input z , $\text{enc}_{\{g(k)\}}(g \cdot z)$ è l'attivazione della feature ottenuta applicando la trasformazione di gruppo g alla feature k per l'input trasformato $g \cdot z$, e la somma è su tutte le coppie (k, g) nella struttura di orbita. Questo termine di regolarizzazione impone l'equivarianza sul

SAE: la risposta della feature $g(k)$ all'input trasformato deve essere uguale alla risposta della feature k all'input originale.

14.3 L'interpretazione geometrica del dizionario SAE nello spazio JEPA

Sia $W_{\text{dec}} = [w_1, w_2, \dots, w_n]$ la matrice dei vettori del dizionario SAE con colonne normalizzate. La decomposizione di una rappresentazione $z \in \mathbb{R}^d$ attraverso il SAE è:

$$z \approx \sum_{\{k : \hat{h}_k > 0\}} \hat{h}_k \cdot w_k + b_{\text{dec}}$$

ossia z è approssimato come combinazione lineare sparsa dei vettori del dizionario con coefficienti $\hat{h}_k > 0$. La *geometria* di questa decomposizione nello spazio JEPA ha un'interpretazione precisa che emerge dalla struttura degli spazi JEPA derivata da LeJEPA (monografia Rapp-JEPA).

LeJEPA dimostra che, per un sistema JEPA addestrato all'optimum, le rappresentazioni degli strati superiori dell'encoder vivono su una *varietà geometrica* determinata dalla struttura del manifold dei dati. Più precisamente, la teoria del collasso rappresentazionale nei sistemi JEPA mostra che la coppia (z_c, z_t) — contesto e target — è distribuita in modo da massimizzare la covarianza tra predizione e target mentre si minimizza la ridondanza (si veda la derivazione dell'ELBO a cinque termini in Var-JEPA). Il risultato geometrico è che lo spazio latente di un JEPA addestrato tende a organizzarsi in modo che le direzioni con alta varianza spiegata nel senso della covarianza context-target corrispondano alle dimensioni più informative per la previsione.

Quando si addestra un SAE su questo spazio, i vettori del dizionario che vengono selezionati dall'algoritmo di sparse coding tendono a corrispondere a queste direzioni di alta covarianza. La decomposizione SAE è quindi una forma di analisi delle componenti principali generalizzata (con una base sovra completa invece di una base ortogonale), orientata a identificare le «direzioni interpretabili» dello spazio latente che contribuiscono maggiormente alla struttura predittiva dell'encoder.

Per rendere questa connessione precisa, si definisca la *varianza spiegata della feature k* nello spazio JEPA come:

$$\sigma^2_k = \text{Cov}(\hat{h}_k, z \cdot w_k)$$

ossia la covarianza tra l'attivazione della feature k e la proiezione della rappresentazione lungo la direzione del vettore del dizionario corrispondente. Le feature con alta σ^2_k sono quelle che catturano direzioni dello spazio JEPA che variano sistematicamente con proprietà predittivamente rilevanti dell'input. Questa metrica è il corrispettivo sparso del contributo alla varianza spiegata nella PCA.

La connessione con la Deep Self-Supervision di V-JEPA 2.1 emerge qui in modo naturale: i quattro livelli intermedi dell'encoder su cui la Deep Self-Supervision calcola la perdita corrispondono a quattro spazi latenti diversi con strutture geometriche diverse. Un SAE addestrato a ciascun livello produrrà un dizionario diverso, con le feature degli strati bassi che corrispondono a proprietà locali (texture, bordi) e le feature degli strati alti che corrispondono a proprietà semantiche (categorie, relazioni spaziali). La traiettoria attraverso i quattro dizionari SAE — come un concetto «radice» si frammenta e riasmonta attraverso la profondità del modello — è il corrispettivo sparse del TCAV gerarchico discusso nel Capitolo 13.

14.4 Applicazione al problema della monosemanticità: evidenza empirica e previsioni teoriche

Il paradigma degli SAE come strumento di interpretabilità meccanicistica è stato sviluppato principalmente nell'ambito dei Large Language Models da Anthropic (Bricken et al. 2023, Templeton et al. 2024), dove gli SAE applicati alle attivazioni dei residual stream dei transformer hanno permesso di identificare feature monosemantiche corrispondenti a concetti linguistici interpretabili (Parigi, capitalismo, errori di programmazione, etc.). Non si dispone ancora di una letteratura equivalente per i sistemi JEPA visivi — questa è una delle lacune che il programma di ricerca tracciato in questo capitolo propone di colmare.

Le previsioni teoriche derivabili dalla struttura degli spazi JEPA biforcati sono le seguenti.

Per lo spazio z_i , ci si attende che il SAE produca feature altamente monosemantiche per proprietà visive locali che siano equivarianti rispetto alle trasformazioni addestrate: la direzione della normale a una superficie, il colore locale di un patch, la frequenza spaziale dominante di una regione. Le feature per proprietà globali (categoria, forma dell'oggetto) saranno invece meno monosemantiche — distribuite su orbite multiple nello spazio di equivarianza — perché l'encoder è ottimizzato per la predittività della prossima vista, non per la classificazione.

Per lo spazio z_{AGG} , ci si attende la distribuzione inversa: feature altamente monosemantiche per proprietà semantiche di alto livello (categoria, funzione, relazioni spaziali tra parti), e feature meno monosemantiche per proprietà locali e dipendenti dalla prospettiva. Questo perché l'aggregazione multi-vista comprime la variabilità trasformativa e lascia emergere soltanto le strutture stabili attraverso le viste.

Un modo per quantificare empiricamente la monosemanticità è il *monosemanticity score* proposto da Bricken et al.: per ciascuna feature k del SAE, si ordina il dataset di input per attivazione $\hat{h}_k$ decrescente e si ispezionano i top-K input che la attivano. Se i top-K input condividono una proprietà comune

interpretabile, la feature è monosemantica. Il monosemanticity score è la proporzione di feature per cui questa proprietà è chiaramente identificabile da osservatori umani.

Nel contesto visivo, questo test è applicabile sia agli encoder JEPA sia alle rappresentazioni dense prodotte da V-JEPA 2.1: i top-K patch (per un SAE applicato alle feature per-patch dello strato finale) che attivano maggiormente la feature k dovrebbero condividere una proprietà visiva interpretabile. Le visualizzazioni PCA di V-JEPA 2.1 mostrate in Mur-Labadia et al. (Figure 14 e 15) — dove i tre componenti principali delle feature-map dell'encoder sono mappati ai canali RGB e mostrano strutture semanticamente coerenti — sono una versione a bassa risoluzione di questo test: la presenza di strutture coerenti nelle mappe PCA è una condizione necessaria ma non sufficiente per la monosemanticità delle feature SAE.

14.5 La connessione tra SAE e identificabilità causale

Vi è una connessione formale tra il problema dell'addestramento di SAE con alta monosemanticità e il problema dell'identificabilità causale discusso nel Capitolo 12. L'IVAE di Khemakhem et al. (2020) dimostra che le variabili latenti di un modello generativo sono causalmente identificabili (cioè univocamente recuperabili modulo una permutazione e una trasformazione per componente) quando si dispone di variabili ausiliarie u che modulano il prior sui latenti. Nel contesto degli SAE, i token di posizione (patch positions) e le azioni (action embeddings) di seq-JEPA fungono da variabili ausiliarie che modulano le distribuzioni delle attivazioni delle feature SAE.

Più precisamente, per un SAE addestrato sullo spazio z_i di seq-JEPA, l'attivazione della feature k per la vista i -esima dipende sia dal contenuto visivo del patch (l'input x_i) sia dalla posizione del patch nell'immagine (la variabile ausiliaria u_i = posizione). L'equazione strutturale causale per la feature k è:

$$\hat{h}_k(x_i) = f_k(\text{contenuto}(x_i), \text{posizione}(x_i)) + \epsilon_k$$

dove f_k è una funzione (approssimata dal SAE) e ϵ_k è rumore residuo. L'identificabilità causale richiede che, per ogni coppia di valori del contenuto (c, c') e di posizione (u, u'), l'equazione $\hat{h}_k(c, u) \neq \hat{h}_k(c', u')$ per almeno un k : le feature SAE devono separare in modo identificabile il contributo del contenuto da quello della posizione.

Questa condizione è soddisfatta per costruzione nel dataset 3DIEBench (dove posizione e contenuto dell'oggetto sono manipolate indipendentemente) e in PLS (Predictive Learning across Saccades) di seq-JEPA (dove la posizione del patch nell'immagine è controllata dal processo di campionamento saccadico). La presenza di Action Conditioning nell'addestramento di seq-JEPA — che conditioning l'encoder sulle trasformazioni tra viste — impone implicitamente la struttura di questo DAG causale nella rete, rendendo le feature SAE più identificabili causalmente.

14.5.1 Prova formale dell'identificabilità della decomposizione SAE sotto action conditioning in seq-JEPA

La sezione 14.5 ha introdotto la connessione tra il problema dell'addestramento di SAE con alta monosemanticità e il problema dell'identificabilità causale, individuando nelle azioni di seq-JEPA il ruolo formale delle variabili ausiliarie del teorema iVAE di Khemakhem et al. (2020). Tuttavia, la verifica che le condizioni del teorema iVAE siano effettivamente soddisfatte dalla struttura di seq-JEPA con action conditioning è stata lasciata implicita. Il presente paragrafo colma questa lacuna, dimostrando che l'action conditioning di seq-JEPA soddisfa le ipotesi di iVAE e quantificando la classe di ambiguità residua della decomposizione SAE.

Richiamo del teorema iVAE. Nel framework iVAE, le osservazioni $x \in \mathbb{R}^d$ sono generate da un modello latente:

$$x = g(z), z_i \sim p(z_i | u) = \exp(T(z_i)^T \lambda(u) - \log Z(u))$$

dove g è il decoder (iniettivo), $u \in \mathbb{R}^k$ è la variabile ausiliaria osservata, $T(z_i) = (T_1(z_i), \dots, T_n(z_i))$ è il vettore delle statistiche sufficienti dell'esponenziale family su z_i , $\lambda(u) = (\lambda_1(u), \dots, \lambda_n(u))$ è il vettore dei parametri naturali condizionati su u , e $Z(u)$ è la funzione di partizione. Il teorema iVAE afferma che, sotto le condizioni:

(iVAE-1) g è diffeomorfo. (iVAE-2) La matrice $\Lambda = [\lambda(u^1) - \lambda(u^0), \lambda(u^2) - \lambda(u^0), \dots, \lambda(u^{\{nk\}} - \lambda(u^0)] \in \mathbb{R}^{\{nk \times nk\}}$ ha rango pieno nk , dove $u^0, u^1, \dots, u^{\{nk\}}$ sono $nk + 1$ valori distinti di u . (iVAE-3) I fattori latenti z_i hanno distribuzioni non-gaussiane e non-degeneri.

allora i fattori latenti z sono identificabili modulo permutazione e trasformazioni componente-per-componente.

Il modello SAE-seq-JEPA come istanza iVAE. Nel contesto di seq-JEPA, le osservazioni sono le rappresentazioni dell'encoder $z_i = E_\theta(x_i) \in \mathbb{R}^d$, e la variabile ausiliaria è l'azione $a_i \in \mathbb{R}^k$ che trasporta la vista i alla vista $i+1$. Le «osservazioni» del SAE sono le attivazioni $\hat{h} = \text{enc_SAE}(z_i) \in \mathbb{R}^n$ (con $n \gg d$ per un dizionario sovra-completo), e l'obiettivo è recuperare i fattori latenti $\hat{h}$ in modo identificabile.

Il modello generativo del SAE-seq-JEPA assume che ogni feature del SAE sia distribuita condizionalmente all'azione secondo una distribuzione esponenziale con statistiche sufficienti dipendenti dall'azione:

$$p(\hat{h}_k | a_i) \propto \exp(T(\hat{h}_k)^T \lambda_k(a_i))$$

dove per una distribuzione spike-and-slab (la prior naturale per feature sparse):

$$T(\hat{h}_k) = (\mathbb{1}[\hat{h}_k > 0], \hat{h}_k \cdot \mathbb{1}[\hat{h}_k > 0])^T \in \mathbb{R}^2$$

$$\lambda_k(a_i) = (\lambda_k^{\text{gate}}(a_i), \lambda_k^{\text{scale}}(a_i))^T \in \mathbb{R}^2$$

dove $\lambda_k^{\text{gate}}(a_i)$ controlla la probabilità che la feature k sia attiva data l'azione a_i , e $\lambda_k^{\text{scale}}(a_i)$ controlla la magnitudine dell'attivazione quando è attiva. La variazione di entrambi i parametri in funzione dell'azione è il meccanismo che impone struttura alla distribuzione delle feature.

Verifica delle condizioni iVAE. Si dimostra ora che le tre condizioni sono soddisfatte dalla struttura di seq-JEPA.

Condizione (iVAE-1): g è diffeomorfo. Il mapping da $\hat{h}$ a z_i attraverso il decoder SAE è $\text{dec_SAE}(\hat{h}) = W_{\text{dec}} \cdot \hat{h} + b_{\text{dec}}$. Poiché il SAE è addestrato con un termine di ricostruzione $\|z_i - \text{dec_SAE}(\text{enc_SAE}(z_i))\|^2 \rightarrow 0$, la composizione $\text{dec_SAE} \circ \text{enc_SAE}$ è approssimativamente l'identità su z_i . Il mapping g nella terminologia iVAE corrisponde a dec_SAE : l'encoder SAE è la sua approssimazione sinistra. La condizione di diffeomorfismo è soddisfatta approssimativamente nella misura in cui la ricostruzione è accurata. Nel limite ideale (ricostruzione perfetta), dec_SAE è iniettivo sullo spazio delle attivazioni sparse.

Condizione (iVAE-2): rango pieno della matrice A . La matrice A deve avere rango $2n$ (poiché le statistiche sufficienti hanno dimensione 2 per feature, e ci sono n feature). La condizione di rango è soddisfatta se le azioni a_i producono variazioni sufficientemente ricche nei parametri naturali $\lambda_k(a_i)$. In seq-JEPA, le azioni corrispondono a trasformazioni rigide dello spazio 3D (rotazioni $SO(3)$ e traslazioni $T(3)$), ovvero a elementi del gruppo $SE(3)$ con sei gradi di libertà. Il numero di parametri naturali per feature è 2 (gate e scale), e il numero di feature è n . Per la condizione di rango pieno, sono necessarie almeno $2n + 1$ azioni distinte in «posizione generale» — ovvero tale che il rango del sistema di equazioni $\sum_{j=1}^{2n+1} p(a^j) \cdot \lambda_k(a^j) = \text{rango pieno}$. Con n tipicamente nel range $\{4096, 8192, 16384\}$ e $SE(3)$ che ha spazio di azione continuo (quindi infinitamente molti valori possibili), la condizione di rango pieno è genericamente soddisfatta per qualsiasi distribuzione di azioni non-degenere. L'unico caso di violazione è se le azioni a_i producono variazioni di λ_k identiche per tutte le feature k — cioè se l'action conditioning non discrimina tra feature diverse — il che contraddirebbe l'utilità stessa del conditioning.

Condizione (iVAE-3): distribuzioni non-gaussiane. La distribuzione spike-and-slab è per definizione non-gaussiana: ha una componente discreta (la probabilità di attivazione) e una componente continua (la magnitudine dell'attivazione). La condizione è soddisfatta strutturalmente.

Teorema (Identificabilità SAE-seq-JEPA). Sotto le condizioni di rango pieno dell'action conditioning di seq-JEPA (condizione iVAE-2 soddisfatta dalla continuità di $SE(3)$), la decomposizione SAE delle rappresentazioni $z_i = E_{\theta}(x_i)$ è identificabile modulo il gruppo di ambiguità $\mathcal{G}_{\text{SAE}} = \mathcal{P}_n \times \mathcal{F}_n$, dove $\mathcal{P}_n$ è il gruppo delle permutazioni delle n feature e $\mathcal{F}_n$ è il gruppo delle trasformazioni invertibili componente per componente sulle attivazioni $\hat{h}_k$. $\square$

La classe di ambiguità residua e la sua interpretazione. Il gruppo $\mathcal{G}_{\text{SAE}}$ $= \mathcal{P}_n \times \mathcal{F}_n$ significa che due soluzioni SAE che producono la stessa distribuzione delle attivazioni condizionate sull'azione differiscono al più per una permutazione delle feature e per trasformazioni monotone componente per componente delle attivazioni. Questo è un risultato di identificabilità molto forte rispetto al caso non-condizionato (senza azioni), in cui la classe di ambiguità residua include l'intero gruppo delle trasformazioni lineari invertibili $\text{GL}(d)$ — qualsiasi rotazione dello spazio latente è indistinguibile.

La contrazione dell'ambiguità da $\text{GL}(d)$ a $\mathcal{G}_{\text{SAE}}$ è il contributo formale dell'action conditioning alla monosemanticità del SAE: senza conditioning, le feature SAE sono identificabili solo modulo rotazione dello spazio latente (ambiguità enorme); con conditioning, le feature sono identificabili modulo permutazione e riscalamento (ambiguità minimale). In pratica, questo significa che il SAE addestrato su z_i con action conditioning produce feature monosemantiche nel senso che le feature corrispondenti a proprietà visive distinte — la direzione normale a una superficie, la frequenza spaziale di una texture, la categoria di un oggetto — sono strettamente separate, non mescolate da rotazioni arbitrarie dello spazio latente. L'action conditioning è pertanto il meccanismo che trasforma il SAE da uno strumento di decomposizione non-interpretabile in uno strumento di interpretabilità strutturata.

Capitolo 15 — RLHF come ottimizzazione vincolata compatibile con gli spazi JEPA

15.1 Il problema dell'allineamento: dalla performance alla conformità ai valori

I capitoli precedenti hanno sviluppato strumenti per *ispezionare* gli spazi latenti dei sistemi JEPA — verificare cosa rappresentano e quanto le loro rappresentazioni siano causalmente strutturate. Ma l'ispezione è soltanto il presupposto dell'allineamento, non l'allineamento stesso. Sapere che il sistema codifica informazione sull'intenzione dell'utente nei propri stati latenti non garantisce che questa informazione sia utilizzata in modo conforme ai valori dell'utente. Per garantire questa conformità è necessario un meccanismo di *ancoraggio normativo* — un processo che modifichi i parametri del sistema in risposta a segnali di preferenza umana.

Il Reinforcement Learning from Human Feedback (RLHF), nella formalizzazione di Stiennon et al. (2020) e Ouyang et al. (2022), è il meccanismo di ancoraggio normativo più maturo e sistematicamente studiato. In sintesi, RLHF aggiunge all'obiettivo di training originale del sistema un termine di reward derivato dalla preferenza umana: un modello di reward $r_\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ — addestrato a predire la preferenza umana tra coppie di output

(y, y') per lo stesso input x — viene usato per ottimizzare il sistema tramite RL verso output con reward più alto.

Nella formulazione standard per i Large Language Models, il problema si scrive come ottimizzazione di:

$$\max_{\pi} \mathbb{E}_{\{x \sim p(x), y \sim \pi(y|x)\}} [r_{\psi}(x, y)] - \beta \cdot D_{\text{KL}}(\pi(y|x) \parallel \pi_{\text{ref}}(y|x))$$

dove π è la policy da ottimizzare, π_{ref} è la policy di riferimento (il modello pre-RLHF), r_{ψ} è il modello di reward, e il termine D_{KL} penalizza le deviazioni troppo grandi dalla policy di riferimento per evitare il reward hacking. Il parametro β bilancia l'ottimizzazione del reward e il mantenimento della vicinanza alla distribuzione di riferimento.

15.2 Formalizzazione di RLHF come problema di ottimizzazione vincolata

La formulazione più rigorosa di RLHF è quella del problema di ottimizzazione vincolata proposta da Ziegler et al. (2019) nella sua versione duale. Invece di aggiungere il termine KL all'obiettivo, si scrive il problema come:

$$\max_{\pi} \mathbb{E}_{\{x \sim p(x), y \sim \pi(y|x)\}} [r_{\psi}(x, y)] \text{ soggetto a: } D_{\text{KL}}(\pi(y|x) \parallel \pi_{\text{ref}}(y|x)) \leq \epsilon \text{ per quasi ogni } x$$

dove ϵ è un budget di deviazione dalla policy di riferimento. Il problema duale di questo problema vincolato è esattamente l'obiettivo RLHF con il termine KL, con $\beta = 1/\lambda$ dove λ è il moltiplicatore di Lagrange del vincolo KL. Questa formulazione rende esplicita la struttura del compromesso: ogni unità di reward guadagnata ha un costo in termini di divergenza dalla distribuzione di riferimento, e β quantifica il tasso di cambio.

La soluzione ottima del problema vincolato ha una forma chiusa (derivata da Peters e Schölkopf 2010, poi generalizzata da Korbak et al. 2022):

$$\pi^*(y|x) = \pi_{\text{ref}}(y|x) \cdot \exp(r_{\psi}(x, y) / \beta) / Z(x)$$

dove $Z(x) = \mathbb{E}_{\{y \sim \pi_{\text{ref}}\}} [\exp(r_{\psi}(x, y) / \beta)]$ è la funzione di partizione. Questa soluzione — una distribuzione di Boltzmann sul reward con temperatura β — è il risultato fondamentale che connette RLHF all'inferenza bayesiana: la policy ottima RLHF è il posteriore bayesiano sulla policy di riferimento quando il reward è interpretato come log-verosimiglianza e β come temperatura.

15.2.1 La soluzione analitica del problema RLHF e le sue implicazioni

La forma chiusa della policy ottima $\pi^*(y | x) = \pi_{\text{ref}}(y | x) \cdot \exp(r(x,y)/\beta) / Z(x)$ ha conseguenze analitiche importanti che meritano di essere svolte esplicitamente. In primo luogo, si noti che il logaritmo del rapporto π^*/π_{ref} è proporzionale al reward:

$$\log \pi^*(y | x) - \log \pi_{\text{ref}}(y | x) = r(x,y)/\beta - \log Z(x)$$

Questa relazione permette di esprimere il reward come funzione della policy ottima e della policy di riferimento — senza richiedere la conoscenza di $Z(x)$ che è in generale intrattabile:

$$r(x,y) = \beta \cdot [\log \pi^*(y | x) - \log \pi_{\text{ref}}(y | x)] + \beta \cdot \log Z(x)$$

Poiché $\log Z(x)$ non dipende da y , nella differenza di reward tra due output y_w e y_l (il setting del RLHF con confronti umani) il termine $\log Z(x)$ si cancella:

$$r(x, y_w) - r(x, y_l) = \beta \cdot [\log \pi^*(y_w | x) - \log \pi_{\text{ref}}(y_w | x)] - \beta \cdot [\log \pi^*(y_l | x) - \log \pi_{\text{ref}}(y_l | x)]$$

Questo è esattamente il nucleo dell'obiettivo DPO di Rafailov et al. (2023): la differenza di reward tra due output è completamente determinata dal rapporto delle probabilità della policy ottima rispetto alla policy di riferimento. La funzione di partizione $Z(x)$ — che sarebbe computazionalmente inaccessibile per uno spazio di output continuo — scompare dall'equazione. Il risultato formale qui presentato chiarisce perché DPO sia equivalente a RLHF con reward model ottimale: entrambi ottimizzano la stessa policy π^* , soltanto attraverso percorsi diversi. RLHF ottimizza prima il reward model r_ψ e poi usa RL; DPO ottimizza direttamente il rapporto π^*/π_{ref} attraverso la loss di Bradley-Terry.

La connessione con gli spazi latenti JEPA emerge quando si sostituisce il condizionamento su x con il condizionamento su $z_{\text{AGG}} = g(E_{\text{JEPA}}(x))$. La policy ottima JEPA-condizionata è:

$$\pi^*_{\text{JEPA}}(y | z_{\text{AGG}}) = \pi_{\text{ref}}(y | z_{\text{AGG}}) \cdot \exp(r(z_{\text{AGG}}, y)/\beta) / Z(z_{\text{AGG}})$$

dove $Z(z_{\text{AGG}}) = \mathbb{E}_{\{y \sim \pi_{\text{ref}}\}}[\exp(r(z_{\text{AGG}}, y)/\beta)]$. Poiché due input x_1 e x_2 con la stessa rappresentazione JEPA $z_{\text{AGG}}(x_1) = z_{\text{AGG}}(x_2)$ ricevono la stessa policy ottima, il fine-tuning RLHF nello spazio latente JEPA è automaticamente **coerente** rispetto all'equivalenza percettiva definita dall'encoder: il sistema produce la stessa risposta per tutti gli input che l'encoder giudica semanticamente equivalenti. Questa coerenza non è garantita dal RLHF standard nello spazio degli input raw, dove piccole perturbazioni dell'input

potrebbero produrre risposte diverse attraverso i meccanismi di reward hacking discussi nella sezione 15.5.

15.3 Compatibilità di RLHF con gli obiettivi di training JEPA

La domanda centrale di questo capitolo è: come si integra il meccanismo RLHF con un sistema il cui backbone percettivo è un encoder JEPA addestrato con l'obiettivo di minimizzazione della distanza coseno nello spazio latente? La risposta dipende da quale componente del sistema si vuole ottimizzare con RLHF.

Vi sono tre modalità principali di integrazione.

La prima modalità — *RLHF sul modulo decisionale a encoder congelato* — mantiene l'encoder JEPA fisso e applica RLHF soltanto ai moduli a valle che producono gli output finali del sistema (per esempio, la policy di controllo, il modulo di linguaggio, il classificatore decisionale). In questo caso, il problema RLHF è:

$$\max_{\pi} \mathbb{E}_{\{x, y \sim \pi(y \mid z_{\text{AGG}}(x))\}} [r_{\psi}(x, y)] - \beta \cdot D_{\text{KL}}(\pi(y \mid z_{\text{AGG}}(x)) \parallel \pi_{\text{ref}}(y \mid z_{\text{AGG}}(x)))$$

dove $z_{\text{AGG}}(x) = g(E_{\text{JEPA}}(x))$ è la rappresentazione aggregata prodotta dall'encoder congelato. La policy π agisce sullo spazio delle rappresentazioni JEPA, non direttamente sull'input raw. Questo approccio ha il vantaggio di preservare le proprietà di invarianza-equivarianza dell'encoder — le proprietà costruttivamente garantite dall'addestramento JEPA — e di limitare il rischio di reward hacking che emerge quando l'encoder è ottimizzato insieme alla policy: se l'encoder potesse cambiare liberamente, potrebbe imparare a trasformare gli input in rappresentazioni che massimizzano il reward del modello r_{ψ} senza mantenere le proprietà di generalizzazione desiderate.

La seconda modalità — *RLHF sulla funzione di reward nello spazio latente* — non usa un modello di reward standard $r_{\psi}(x, y)$, ma definisce il reward direttamente nello spazio latente JEPA:

$$r_{\text{JEPA}}(z_{\text{AGG}}, y) = r_{\psi}(z_{\text{AGG}}, y)$$

dove il modello di reward prende come input la rappresentazione aggregata anziché l'input raw. Questo approccio sfrutta l'astrazione già compiuta dall'encoder JEPA: due input x e x' che hanno la stessa rappresentazione aggregata z_{AGG} — perché l'encoder ha compresso le variazioni superficiali irrilevanti — ricevono lo stesso reward. Questo riduce la variabilità del segnale di reward e facilita l'addestramento del modello r_{ψ} .

La terza modalità — *RLHF con fine-tuning congiunto* — ottimizza congiuntamente l'encoder JEPA e la policy attraverso il gradiente del reward. Questa modalità è la più potente ma anche la più rischiosa: consente all'encoder di adattarsi al segnale di preferenza, ma al costo di potenzialmente degradare le proprietà

costruttivamente garantite dall'addestramento JEPA se il segnale di reward è distorto o limitato.

15.4 Il modello di reward nello spazio latente JEPA e la derivazione del suo gradiente

Per la seconda modalità di integrazione — RLHF con reward nello spazio latente — si deriva qui il gradiente del training objective rispetto ai parametri dell'encoder.

Sia $r_\psi(z_AGG, y) = \varrho_\psi(z_AGG) \cdot \varphi_\psi(y)$ il modello di reward fattorizzato in un termine dipendente dalla rappresentazione e un termine dipendente dall'output (questa fattorizzazione è una semplificazione utile per la derivazione, non una restrizione necessaria). La derivata del reward rispetto ai parametri dell'encoder θ (che compaiono in $z_AGG = g(E_\theta(x))$) è:

$$\partial r_\psi / \partial \theta = (\partial r_\psi / \partial z_AGG) \cdot (\partial z_AGG / \partial z_i) \cdot (\partial z_i / \partial \theta) \cdot \varphi_\psi(y)$$

dove $(\partial r_\psi / \partial z_AGG) \in \mathbb{R}^{1 \times d}$ è il gradiente del reward rispetto alla rappresentazione aggregata, $(\partial z_AGG / \partial z_i) \in \mathbb{R}^{d \times d}$ è la jacobiana dell'aggregatore rispetto alle rappresentazioni dell'encoder (matrice $d \times d$), e $(\partial z_i / \partial \theta)$ è la jacobiana dell'encoder rispetto ai suoi parametri (matrice $d \times |\theta|$).

In un transformer aggregatore come quello di seq-JEPA, la jacobiana $(\partial z_AGG / \partial z_i)$ non è diagonale: il token AGG è prodotto attraverso l'attenzione multi-testa su tutti i token di input (z_j, a_j) per $j = 1, \dots, M$, e il gradiente di z_AGG rispetto a z_i tiene conto dell'effetto indiretto attraverso le modifiche ai pesi di attenzione prodotte dal cambiamento di z_i . Questo gradiente è computabile attraverso backpropagation standard, ma la sua interpretazione è complessa: una piccola perturbazione di z_i influenza z_AGG non soltanto attraverso la propria contribuzione diretta, ma attraverso le modifiche ai pesi di attenzione che reindirizzano l'elaborazione degli altri token.

Nel regime in cui l'encoder è congelato (prima modalità), questo problema non si pone: si ottimizza soltanto la policy a valle, e il gradiente dell'encoder non è mai calcolato.

15.5 Il problema del reward hacking negli spazi latenti JEPA

Il *reward hacking* — la scoperta da parte del sistema di strategie che massimizzano il reward senza soddisfare l'intenzione dell'annotatore umano — è il problema centrale dell'RLHF. Nella sua forma più classica, il reward hacking emerge quando il modello di reward r_ψ è addestrato su un numero limitato di confronti umani e il sistema ottimizzato con RL trova percorsi nello spazio degli output che massimizzano r_ψ senza corrispondere alle preferenze che gli annotatori umani avrebbero espresso su quegli output.

Negli spazi latenti JEPA, il reward hacking può manifestarsi in modi specifici alla struttura di questi spazi. Per la prima modalità di integrazione (encoder congelato), il reward hacking è limitato allo spazio z_AGG : il sistema cerca di produrre rappresentazioni aggregate con proprietà che il modello di reward premia, senza modificare l'encoder. Poiché l'encoder JEPA è addestrato a essere invariante alle variazioni irrilevanti, questa forma di reward hacking è più difficile che in sistemi con encoder adattivi: le rappresentazioni z_AGG sono vincolate a riflettere la struttura dell'input filtrata dall'encoder JEPA, e le «scorciatoie» tipiche del reward hacking (variazioni di stile superficiali, artefatti di frequenza, correlazioni spurie del dataset) sono già state compresse dall'encoder.

Per la terza modalità (fine-tuning congiunto), il reward hacking è potenzialmente più grave: l'encoder potrebbe imparare a trasformare gli input in rappresentazioni che massimizzano il modello di reward senza preservare le proprietà di invarianza-equivarianza. La penalità KL rispetto alla policy di riferimento π_ref — il secondo termine dell'obiettivo RLHF — funziona come un vincolo di regolarizzazione, ma per i sistemi JEPA si potrebbe aggiungere una penalità più diretta sulle proprietà dell'encoder:

$$\mathcal{L}_{RLHF_reg}(\theta) = -\mathbb{E}[r_\psi(z_AGG, y)] + \beta \cdot D_{KL}(\pi \parallel \pi_ref) + \gamma \cdot \mathcal{L}_{JEPA_finetuning}(\theta)$$

dove $\mathcal{L}_{JEPA_finetuning}(\theta)$ è una versione leggera dell'obiettivo JEPA originale — la perdita di predizione su un piccolo dataset di «replay» degli esempi di training originali — che mantiene le proprietà dell'encoder durante il fine-tuning. Questa idea — chiamata *continual pretraining* o *elastic weight consolidation* nella letteratura del continual learning — è il modo più diretto per prevenire la degradazione delle proprietà JEPA durante l'ottimizzazione con RLHF.

15.6 Il Direct Preference Optimization come alternativa senza RL

Il Direct Preference Optimization (DPO) di Rafailov et al. (2023) è un'alternativa a RLHF che elimina la necessità di addestrare esplicitamente un modello di reward e di usare algoritmi RL, riformulando l'ottimizzazione delle preferenze come un problema di maximum likelihood su dati di confronto.

La derivazione di DPO parte dall'osservazione che la soluzione ottima del problema RLHF ha la forma $\pi^*(y|x) = \pi_ref(y|x) \cdot \exp(r(x,y) / \beta) / Z(x)$. Invertendo questa relazione, si ottiene:

$$r^*(x, y) = \beta \cdot \log(\pi^*(y|x) / \pi_ref(y|x)) + \beta \cdot \log Z(x)$$

La preferenza umana ($y_w > y_l \mid x$) — y_w è preferito a y_l dato x — è modellata da un modello di Bradley-Terry:

$$p(y_w > y_l \mid x) = \sigma(r^*(x, y_w) - r^*(x, y_l)) = \sigma(\beta \cdot \log(\pi^*(y_w \mid x) / \pi_{\text{ref}}(y_w \mid x)) - \beta \cdot \log(\pi^*(y_l \mid x) / \pi_{\text{ref}}(y_l \mid x)))$$

dove σ è la funzione sigmoide e $Z(x)$ si cancella nella differenza. Massimizzando la verosimiglianza del modello di preferenza sul dataset di confronti $D = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N$ rispetto a π^* , si ottiene:

$$\mathcal{L}_{\text{DPO}}(\pi^*; \pi_{\text{ref}}) = -\mathbb{E}_{\{(x, y_w, y_l) \sim D\}} [\log \sigma(\beta \cdot \log(\pi^*(y_w \mid x) / \pi_{\text{ref}}(y_w \mid x)) - \beta \cdot \log(\pi^*(y_l \mid x) / \pi_{\text{ref}}(y_l \mid x)))]$$

L'ottimizzazione di questa perdita rispetto ai parametri di π^* non richiede campionamento RL, rende il gradiente direttamente calcolabile, e garantisce implicitamente il vincolo KL attraverso la struttura del rapporto π^*/π_{ref} .

Nel contesto dei sistemi JEPA, DPO è particolarmente attraente per la prima modalità di integrazione (encoder congelato): invece di addestrare un modello di reward separato $r_\psi(z_{\text{AGG}}, y)$ e poi fare RL, si ottimizza direttamente la policy $\pi(y \mid z_{\text{AGG}})$ attraverso DPO, usando le rappresentazioni z_{AGG} come input al posto degli input raw. La perdita DPO applicata allo spazio z_{AGG} diventa:

$$\mathcal{L}_{\text{DPO_JEPA}}(\pi; \pi_{\text{ref}}, E_{\text{JEPA}}) = -\mathbb{E}_{\{(x, y_w, y_l) \sim D\}} [\log \sigma(\beta \cdot \log(\pi(y_w \mid z_{\text{AGG}}(x)) / \pi_{\text{ref}}(y_w \mid z_{\text{AGG}}(x))) - \beta \cdot \log(\pi(y_l \mid z_{\text{AGG}}(x)) / \pi_{\text{ref}}(y_l \mid z_{\text{AGG}}(x))))]$$

dove $z_{\text{AGG}}(x) = g(E_{\text{JEPA}}(x))$ con E_{JEPA} congelato. La differenza rispetto al DPO standard è che il condizionamento avviene sulle rappresentazioni JEPA anziché sugli input raw: due input che l'encoder JEPA giudica equivalenti (stessa z_{AGG}) ricevono la stessa policy, il che garantisce una forma di coerenza decisionale rispetto all'equivalenza percettiva JEPA.

15.7 La tensione fondamentale tra ottimizzazione delle preferenze e struttura di equivarianza

Un'osservazione finale, cruciale per la comprensione dei limiti della compatibilità tra RLHF e il paradigma JEPA. Il training objective di RLHF spinge il sistema verso output con alto reward, il che può entrare in tensione con la struttura di equivarianza che l'encoder JEPA ha appreso durante il preaddestramento.

Consideriamo un esempio concreto nel dominio della robotica: un sistema di manipolazione con encoder JEPA addestrato ad essere equivariante rispetto alle rotazioni degli oggetti. Il modello di reward umano potrebbe preferire approcci di presa da certe prospettive — non perché quelle prospettive siano causalmente migliori per il successo della presa, ma perché l'annotatore umano ha familiarità con certi pattern visivi. Se il fine-tuning congiunto modifica l'encoder per privilegiare quelle prospettive, si degrada l'equivarianza — la

capacità del sistema di riconoscere che lo stesso oggetto, visto da angoli diversi, richiede la stessa politica di presa.

Questa tensione non è risolvibile con semplici vincoli di regolarizzazione: è una tensione strutturale tra due obiettivi diversi — l'ottimizzazione delle preferenze come le vede un valutatore umano limitato, e la codifica della struttura causale del mondo come la esprime l'equivarianza rispetto alle trasformazioni fisiche. Il modo più promettente per riconciliarli è l'uso di *preference data che rispettano l'equivarianza*: raccogliere giudizi di preferenza su coppie di output che differiscono per proprietà rilevanti per il task (qualità della presa, precisione della deposizione) mentre si neutralizzano le variazioni trasformazionali (mostrando le stesse prese da diverse prospettive e chiedendo giudizi sulla qualità indipendentemente dalla prospettiva). Questo tipo di data collection epistemicamente informata è il presupposto per un RLHF che non degradi la struttura causale appresa dal preaddestramento JEPA.

15.7.1 Bound formale sulla degradazione dell'equivarianza dell'encoder JEPA sotto ottimizzazione RLHF: la metrica `sim_CAV` come indicatore

La sezione 15.7 ha descritto la tensione tra l'ottimizzazione delle preferenze umane (RLHF) e la struttura di equivarianza dell'encoder JEPA, concludendo che la raccolta di preference data equivariante è il modo più promettente per riconciliarla. Questa conclusione, corretta nella sostanza, manca di una formalizzazione che renda la tensione misurabile e il processo di degradazione quantificabile. Il presente paragrafo deriva un bound sulla velocità di degradazione della metrica `sim_CAV` — già introdotta nella sezione 13.2 — in funzione del numero di step di gradient ascent sull'obiettivo RLHF, utilizzando la struttura della jacobiana dell'encoder rispetto ai parametri di policy.

La metrica `sim_CAV` come misura dell'equivarianza. Sia $\text{sim_CAV}(C, E) = |\mathbf{v}_C^{\{\text{rot}=0^\circ\}} \cdot \mathbf{v}_C^{\{\text{rot}=90^\circ\}}| / (\|\mathbf{v}_C^{\{\text{rot}=0^\circ\}}\| \cdot \|\mathbf{v}_C^{\{\text{rot}=90^\circ\}}\|)$ la similarità coseno tra i CAV del concetto C calcolati su esempi con prospettiva frontale e su esempi ruotati di 90° , come definita nella sezione 13.2. Valori di `sim_CAV` vicini a 0 indicano alta equivarianza (le prospettive diverse producono direzioni concettuali diverse, come atteso per uno spazio equivariante). Valori vicini a 1 indicano collasso dell'equivarianza (le prospettive diverse producono la stessa direzione — l'encoder non distingue più la prospettiva). Nella monografia si usa la convenzione che valori di $\text{sim_CAV} \approx 0$ su z_i corrispondano alla struttura desiderata dell'encoder equivariante.

Il modello di fine-tuning JEPA con RLHF. Si consideri il fine-tuning congiunto dell'encoder (la terza modalità di integrazione descritta nella sezione 15.3), in cui i parametri dell'encoder $\theta \in \mathbb{R}^{\{|\theta|\}}$ vengono aggiornati con gradient ascent sul reward RLHF:

$$\theta^{\{(t+1)\}} = \theta^{\{(t)\}} + \eta \cdot \partial \mathcal{L}_{\text{RLHF}}(\theta^{\{(t)\}}) / \partial \theta$$

dove η è il tasso di apprendimento e $\mathcal{L}_{\text{RLHF}} = \mathbb{E}[\mathbf{r}_\psi(\mathbf{z}_{\text{AGG}}, \mathbf{y})] - \beta \cdot D_{\text{KL}}(\pi \parallel \pi_{\text{ref}})$ è l'obiettivo RLHF dalla sezione 15.5. Dopo t passi di aggiornamento, i parametri dell'encoder si trovano in $\theta^{\{t\}} = \theta^{\{0\}} + \eta \cdot \sum_{s=0}^{t-1} \nabla_{\theta} \mathcal{L}_{\text{RLHF}}(\theta^{\{s\}})$.

La propagazione del cambiamento dei parametri sull'equivarianza. Il CAV del concetto C a prospettiva rot è il vettore normale dell'SVM addestrato su $\{E_{\theta}(x)\}_{x \in D_C^{\{\text{rot}\}} \cup D_{C'}^{\{\text{rot}\}}\}$. La variazione del CAV dovuta a una variazione $\delta\theta$ dei parametri dell'encoder è:

$$\delta v_C^{\{\text{rot}\}} = (\partial v_C^{\{\text{rot}\}} / \partial \theta) \cdot \delta \theta$$

La derivata $\partial v_C^{\{\text{rot}\}} / \partial \theta$ è la jacobiana del CAV rispetto ai parametri dell'encoder. Per la teoria della perturbazione degli SVM a margine massimo (Kumar e Turney, 2012), questa jacobiana è:

$$\partial v_C^{\{\text{rot}\}} / \partial \theta \approx \sum_{i \in \text{SV}} \alpha_i^{\{\text{rot}\}} \cdot y_i \cdot (\partial E_{\theta}(x_i) / \partial \theta)$$

dove la somma è sugli indici i dei vettori di supporto dell'SVM, $\alpha_i^{\{\text{rot}\}}$ sono i moltiplicatori di Lagrange ottimali, $y_i \in \{+1, -1\}$ sono le etichette, e $\partial E_{\theta}(x_i) / \partial \theta \in \mathbb{R}^{d \times |\theta|}$ è la jacobiana dell'encoder rispetto ai parametri.

La variazione del coseno sim_CAV in funzione di $\delta\theta$ è, al primo ordine:

$$\delta \text{sim_CAV} = (\partial \text{sim_CAV} / \partial v_C^{\{\text{rot}=0^\circ\}}) \cdot \delta v_C^{\{\text{rot}=0^\circ\}} + (\partial \text{sim_CAV} / \partial v_C^{\{\text{rot}=90^\circ\}}) \cdot \delta v_C^{\{\text{rot}=90^\circ\}}$$

Usando la formula del coseno e le sue derivate rispetto ai vettori:

$$\partial(v \cdot w / (\|v\| \|w\|)) / \partial v = (w - (v \cdot w / \|v\|^2) v) / (\|v\| \cdot \|w\|) = (w - \cos_angle \cdot v_{\text{norm}}) / \|v\|$$

e simmetricamente per $\partial/\partial w$. Nel punto di partenza $\text{sim_CAV} \approx 0$ (encoder equivariante preaddestrato), il termine di primo ordine si semplifica:

$$\delta \text{sim_CAV} \approx [\delta v_C^{\{\text{rot}=90^\circ\}} \cdot v_C^{\{\text{rot}=0^\circ\}} + \delta v_C^{\{\text{rot}=0^\circ\}} \cdot v_C^{\{\text{rot}=90^\circ\}}] / (\|v_C^{\{\text{rot}=0^\circ\}}\| \cdot \|v_C^{\{\text{rot}=90^\circ\}}\|)$$

Teorema (Bound di degradazione dell'equivarianza). Sotto le assunzioni che (i) i vettori di supporto degli SVM non cambino nel corso del fine-tuning (approssimazione di primo ordine), (ii) il gradiente RLHF $\nabla_{\theta} \mathcal{L}_{\text{RLHF}}$ ha norma limitata da G , (iii) l'encoder ha costante di Lipschitz L_E ($\|\partial E_{\theta}(x)/\partial \theta\|_F \leq L_E$), il bound sulla degradazione di sim_CAV dopo t passi di aggiornamento è:

$$|\text{sim_CAV}^{\{t\}} - \text{sim_CAV}^{\{0\}}| \leq 2 \cdot \eta \cdot t \cdot G \cdot L_E \cdot |\text{SV}| \cdot \max_i \alpha_i$$

dove $|SV|$ è il numero di vettori di supporto e α_i sono i moltiplicatori ottimali dell'SVM.

La dimostrazione è immediata dalla composizione delle disuguaglianze triangolari sulle singole variazioni $\delta \text{sim_CAV}^{\{s\}}$ a ogni passo $s = 0, \dots, t-1$, usando che ogni variazione è maggiorata da $\|\delta \theta^{\{s\}}\| \cdot \|\partial v_C^{\{\text{rot}\}} / \partial \theta\| \leq \eta G \cdot L_E \cdot |SV| \cdot \max_i \alpha_i$. $\square$

Interpretazione del bound e implicazioni operative. Il bound cresce linearmente nel numero di step di fine-tuning t , nel tasso di apprendimento η , e nella norma del gradiente RLHF G , ed è amplificato dalla costante di Lipschitz dell'encoder L_E e dalla complessità del margine dell'SVM ($|SV| \cdot \max_i \alpha_i$). Ciascuno di questi fattori suggerisce una strategia di mitigazione della degradazione:

Ridurre η — tassi di apprendimento piccoli nel fine-tuning RLHF preservano l'equivarianza più a lungo. Questo giustifica l'uso di LoRA (Low-Rank Adaptation) con rank basso nei layer dell'encoder durante il fine-tuning RLHF: il rank basso della matrice di aggiornamento $\Delta W = BA$ implica che la jacobiana effettiva dell'encoder rispetto ai parametri LoRA ha rango al più r (il rank di LoRA), riducendo L_E di un fattore $\sqrt{r} / \sqrt{d}$.

Ridurre G — la penalità KL rispetto alla policy di riferimento (il secondo termine di $\mathcal{L}_{\text{RLHF}}$ con coefficiente β) limita la norma del gradiente RLHF: un β elevato riduce G e quindi rallenta la degradazione dell'equivarianza. Questa è una seconda giustificazione, oltre alla prevenzione del reward hacking, per mantenere β elevato nelle fasi iniziali del fine-tuning.

Ridurre L_E — encoder con jacobiana a norma piccola (come i sistemi JEPA con obiettivo di soppressione delle variazioni irrilevanti, discussi nella sezione 8.3.1) degradano più lentamente sotto RLHF. La proprietà di jacobiana compressa che protegge dal sim-to-real gap protegge simultaneamente dall'erosione dell'equivarianza sotto ottimizzazione delle preferenze: entrambe le garanzie derivano dalla stessa proprietà strutturale dell'encoder JEPA.

Aggiungere la penalità di equivarianza $\mathcal{L}_{\text{JEPA_finetuning}}$ — il termine di continual pretraining proposto nella sezione 15.5 agisce direttamente su sim_CAV : mantiene l'obiettivo JEPA attivo durante il fine-tuning, il che equivale a un vincolo soft $\text{sim_CAV} \approx 0$ nel landscape dell'ottimizzazione. Il coefficiente γ di questo termine calibra il trade-off tra velocità di ottimizzazione del reward (preferisce γ piccolo) e preservazione dell'equivarianza (preferisce γ grande). Il bound del teorema fornisce la calibrazione quantitativa: γ dovrebbe essere scelto in modo che la velocità di degradazione γ -indotta dell'obiettivo RLHF sia comparabile alla velocità di degradazione $\eta G L_E |SV| \max_i \alpha_i$ dell'equivarianza, bilanciando i due termini nel landscape complessivo.

Il bound quantifica in modo preciso una delle tensioni fondamentali del paradigma JEPA applicato all'allineamento: l'ottimizzazione delle preferenze

umane, mediata da RLHF, erode strutturalmente l'equivarianza che JEPA ha acquisito durante il preaddestramento. La velocità di erosione è controllabile tramite i quattro lever operativi derivati dal bound, e la sua misura è monitorabile in tempo reale attraverso il calcolo periodico di `sim_CAV` su un dataset di riferimento fisso — un segnale di early-stopping per il fine-tuning RLHF che segnala quando l'equivarianza si è degradata oltre una soglia accettabile.

Capitolo 16 — Scalable oversight come problema fondamentale di allineamento dei sistemi AMI

16.1 Il nodo del superallineamento: quando il supervisore non è più sufficiente

Il paradigma RLHF trattato nel Capitolo 15 assume implicitamente che il supervisore umano sia in grado di valutare la qualità degli output del sistema in modo affidabile. Questa assunzione, ragionevole per i sistemi correnti che operano su task di scrittura, traduzione o risposta a domande generali, diventa progressivamente insostenibile man mano che le capacità dei sistemi aumentano. Un sistema AMI che opera al livello di un esperto medico specialistico, produce prove matematiche avanzate, o progetta esperimenti biologici non può essere supervisionato da annotatori generalisti — e nemmeno da esperti singoli, che per definizione mancano della competenza trasversale necessaria a valutare output che integrano più domini specialistici.

Questo problema ha un nome preciso nella letteratura dell'allineamento: *scalable oversight* (Amodei et al. 2016). La sfida è: come possiamo garantire l'allineamento di sistemi che sono più competenti dei loro supervisori umani? Bowman et al. (2022) hanno formalizzato il problema distinguendo due regimi: il regime *sub-umano* (il sistema è meno capace del supervisore su tutti i task rilevanti) e il regime *sovraumano* (il sistema supera il supervisore su almeno alcuni task rilevanti). RLHF standard è efficace nel regime sub-umano, ma il suo segnale di supervisione si degrada nel regime sovraumano per ragioni sia pratiche (l'annotatore non sa distinguere output corretti da output plausibili ma errati) sia teoriche (un modello di reward addestrato su preferenze umane subottimali è un proxy rumoroso e distorto della qualità reale).

Diverse strategie sono state proposte per affrontare questo problema. La *weak-to-strong generalization* di Burns et al. (2024) studia se un modello forte può superare i propri supervisori deboli estraendo conoscenza che il modello già possiede implicitamente dai propri parametri — una forma di auto-supervisione guidata da un supervisore imperfetto. Il *debate* di Khan et al.

(2024) usa argomenti avversariali tra copie del sistema per rendere verificabili da un giudice umano le affermazioni di correttezza — sfruttando il fatto che verificare un argomento è spesso più facile che generarlo. La *recursive task decomposition* di Wu et al. (2021) decompone task complessi in sotto-task elementari su cui gli umani possono esercitare supervisione diretta.

Yin, Ishida e Sugiyama (2026) introducono un quarto approccio — la *supervisione umana partizionata attraverso etichette complementari* — che occupa una nicchia specifica e ortogonale rispetto alle strategie precedenti. L'intuizione fondamentale è che anche nelle discipline più specialistiche, gli esperti possiedono conoscenza *complementare*: pur non sapendo qual è la risposta corretta in un dominio diverso dal proprio, sanno con alta affidabilità qual è la risposta sbagliata. Un cardiologo non sa diagnosticare un tumore raro, ma sa affermare con certezza «questo non è un problema cardiovascolare». Un analista del settore immobiliare non sa classificare correttamente un'azienda nel settore farmaceutico, ma sa escludere la classificazione nel settore immobiliare.

16.2 Il framework delle etichette complementari: formalizzazione

Il setting formale di Yin et al. è il seguente. Sia un task di valutazione a scelta multipla con K opzioni per item. Ogni item è rappresentato da un input $x \in \mathcal{X}$ con un'etichetta ordinaria $y \in \{1, \dots, K\}$ — l'indice dell'opzione corretta. Un sistema AI $f : \mathcal{X} \rightarrow \{1, \dots, K\}$ produce una previsione $\hat{y} = f(x)$. L'*accuratezza top-1* è $A = \Pr(f(X) = Y)$ dove la probabilità è rispetto alla distribuzione congiunta D su (X, Y) .

Un'etichetta complementare $\bar{Y} \neq Y$ è un indice campionato dalla distribuzione uniforme sui $K-1$ indici non-corretti:

$$p(\bar{Y} = k \mid Y) = 1/(K-1) \text{ per ogni } k \neq Y$$

Questa distribuzione uniforme è garantita dal protocollo di raccolta: per ogni item, si sceglie casualmente un annotatore esperto nel dominio dell'opzione k (tra i K esperti, uno per dominio), e si chiede «l'opzione k è la risposta corretta?». Se risponde «sì», si registra un'etichetta ordinaria con valore k . Se risponde «no», si registra un'etichetta complementare con valore k . Poiché ogni annotatore viene scelto con uguale probabilità $1/K$ e non ha pregiudizi sistematici verso certi indici (ogni esperto è responsabile soltanto del proprio dominio), la distribuzione delle etichette complementari è uniforme sui non-corretti per costruzione.

16.2.1 Proprietà statistiche della variabile di successo W e della sua distribuzione

La variabile indicatrice $W = \mathbb{1}[\hat{y} \neq \bar{Y}]$ è una variabile di Bernoulli con parametro $q = P(\hat{y} \neq \bar{Y})$. La derivazione del legame tra q e l'accuratezza $A =$

$P(\hat{y} = Y)$ procede decomponendo la probabilità su due casi mutuamente esclusivi: il sistema predice correttamente ($\hat{y} = Y$, con probabilità A) e il sistema predice incorrettamente ($\hat{y} \neq Y$, con probabilità $1-A$).

Nel caso $\hat{y} = Y$: poiché $Y \neq \bar{Y}$ per costruzione (l'etichetta complementare è diversa dall'etichetta corretta), si ha certamente $\hat{y} \neq \bar{Y}$, dunque $P(W = 1 \mid \hat{y} = Y) = 1$.

Nel caso $\hat{y} \neq Y$: il sistema ha predetto un'opzione sbagliata $k \neq Y$. L'etichetta complementare $\bar{Y}$ è campionata uniformemente da $\{1, \dots, K\} \setminus \{Y\}$. La probabilità che $W = 1$, ossia che $\hat{y} \neq \bar{Y}$, è la probabilità che l'etichetta complementare non coincida con la predizione sbagliata:

$$P(W = 1 \mid \hat{y} \neq Y) = P(\bar{Y} \neq \hat{y} \mid \hat{y} \neq Y) = (K-2)/(K-1)$$

poiché $\bar{Y}$ è uniforme su $K-1$ valori e uno di essi ($\hat{y}$ stesso) produce $W = 0$. Combinando i due casi con la legge della probabilità totale:

$$q = P(W = 1) = A \cdot 1 + (1-A) \cdot (K-2)/(K-1) = A + (1-A)(K-2)/(K-1)$$

$$q = [A(K-1) + (1-A)(K-2)] / (K-1) = [A + K - 2] / (K-1)$$

Invertendo questa relazione si ottiene lo stimatore non distorto:

$$A = (K-1) \cdot q - (K-2)$$

La non-distorsione dello stimatore $\hat{A}_{\text{comp}} = (K-1) \cdot \hat{q} - (K-2)$ segue immediatamente dalla linearità dell'aspettativa: $\mathbb{E}[\hat{A}_{\text{comp}}] = (K-1) \cdot \mathbb{E}[\hat{q}] - (K-2) = (K-1) \cdot q - (K-2) = A$, come richiesto. La varianza è:

$$\text{Var}(\hat{A}_{\text{comp}}) = (K-1)^2 \cdot \text{Var}(\hat{q}) = (K-1)^2 \cdot q(1-q)/n_c$$

dove si è usato il fatto che le n_c variabili W_i sono i.i.d. Bernoulli(q) per il protocollo di selezione uniforme degli annotatori. Sostituendo $q = (A+K-2)/(K-1)$:

$$\begin{aligned} \text{Var}(\hat{A}_{\text{comp}}) &= (K-1)^2 \cdot [(A+K-2)/(K-1)] \cdot [1-(A+K-2)/(K-1)] / n_c \\ &= (A+K-2)(1-A) \cdot (K-1) / [(K-1)^2 \cdot 1/(K-1)] / n_c \\ &= (A+K-2)(1-A) / n_c \end{aligned}$$

Questa espressione mostra che la varianza dello stimatore complementare cresce con K — il numero di opzioni — e decresce con n_c — il numero di etichette complementari. Il confronto con la varianza dello stimatore ordinario $\text{Var}(\hat{A}_{\text{ord}}) = A(1-A)/n_o$ mostra che il rapporto è $\text{Var}(\hat{A}_{\text{comp}})/\text{Var}(\hat{A}_{\text{ord}}) = (A+K-2) \cdot n_o / (A \cdot n_c)$, che per $n_c = n_o$ vale $(A+K-2)/A > 1$ per qualsiasi $K \geq 3$: ogni etichetta complementare porta meno informazione di una etichetta ordinaria, ma il loro costo di raccolta è sistematicamente inferiore quando l'expertise richiesta per fornirne un giudizio definitivo è inaccessibile.

La variabile ausiliaria di successo del sistema è:

$$W = \mathbb{1}[\hat{y} \neq \bar{Y}] \in \{0, 1\}$$

ossia $W = 1$ se il sistema non ha scelto l'opzione complementare. L'attesa di W è:

$$\mathbb{E}[W] = \Pr(f(X) \neq \bar{Y}) = \Pr(f(X) = Y) \cdot 1 + \Pr(f(X) \neq Y) \cdot (1 - 1/(K-1)) = A + (1-A) \cdot (K-2)/(K-1) = (A + K - 2)/(K - 1)$$

che si può abbreviare come $q = (A + K - 2)/(K - 1)$. Invertendo questa relazione, si ottiene lo *stimatore complementare*:

$$\hat{A}_{\text{comp}} = (K - 1) \cdot \hat{q} - (K - 2)$$

dove $\hat{q} = (1/n_c) \sum_{i=1}^{n_c} W_i$ è la media campionaria. La correttezza (unbiasedness) di questo stimatore segue direttamente: $\mathbb{E}[\hat{A}_{\text{comp}}] = (K-1) \cdot \mathbb{E}[\hat{q}] - (K-2) = (K-1) \cdot q - (K-2) = A$. La derivazione esplicita procede come dimostrato in Appendice A del paper di Yin et al., che identifica questo risultato come la specializzazione alla perdita 0-1 dell'identità di riscrittura del rischio complementare di Ishida et al. (2019).

16.3 Analisi della varianza e il numero di etichette complementari necessarie

Uno stimatore non distorto non è di per sé sufficiente per l'uso pratico: è necessario che la sua varianza sia controllabile. La varianza dello stimatore complementare è derivabile dall'espressione di W come variabile di Bernoulli con parametro q :

$$\text{Var}(\hat{A}_{\text{comp}}) = (K-1)^2 \cdot \text{Var}(\hat{q}) = (K-1)^2 \cdot q(1-q)/n_c = (A+K-2)(1-A)/n_c$$

Confrontando con la varianza dello stimatore ordinario $\hat{A}_{\text{ord}} = S_o/n_o$ (dove $S_o = \sum_j z_j$ è il numero di previsioni corrette e $z_j = \mathbb{1}[\hat{y}_j = y_j]$):

$$\text{Var}(\hat{A}_{\text{ord}}) = A(1-A)/n_o$$

Il rapporto di varianza è:

$$\text{Var}(\hat{A}_{\text{comp}}) / \text{Var}(\hat{A}_{\text{ord}}) = (A + K - 2) \cdot n_o / (A \cdot n_c)$$

che è sempre maggiore di 1 per $K \geq 3$ e $n_c = n_o$, confermando che le etichette complementari sono «più deboli» delle ordinarie in senso statistico: ciascuna porta meno informazione sull'accuratezza vera. Per eguagliare la varianza dello stimatore ordinario con n_o etichette ordinarie, il numero di etichette complementari necessarie è:

$$n_c = (1 + (K-2)/A) \cdot n_o$$

Questo risultato ha un'interpretazione intuitiva: quando A è grande (il sistema è molto accurato), servono relativamente poche etichette complementari aggiuntive per compensare la loro debolezza individuale; quando A è piccolo, ne servono molte di più perché la maggior parte delle etichette complementari porta scarsa informazione (il sistema evita spesso l'opzione complementare anche quando sbaglia semplicemente perché ci sono $K-1$ opzioni da sbagliare). Per $K = 4$ e $A = 0.5$, il numero di etichette complementari necessarie è $(1 + 3/0.5) \cdot n_o = 7 \cdot n_o$ — per ogni etichetta ordinaria che sostituisce, servono 7 etichette complementari.

16.3.1 Il numero ottimo di opzioni K : un'analisi dell'efficienza informativa attraverso la struttura della varianza

La sezione 16.3 ha ricavato che per eguagliare la varianza dello stimatore ordinario con n_o etichette ordinarie servono $n_c = (1 + (K-2)/A) \cdot n_o$ etichette complementari. Questa formula mostra che il rapporto di efficienza delle etichette complementari rispetto a quelle ordinarie dipende criticamente da K — il numero di opzioni del task. Manca tuttavia un'analisi dell'effetto di K non soltanto come parametro esogeno (il numero di opzioni è fissato dal task), ma come variabile di progetto che il ricercatore può ottimizzare nella costruzione del benchmark di valutazione. La presente sottosezione colma questa lacuna derivando il numero ottimale di opzioni K^* che massimizza l'efficienza informativa delle etichette complementari per un dato livello di accuratezza del sistema A e un dato costo relativo $c_r = c_c / c_o$ del costo di raccolta di un'etichetta complementare rispetto a una ordinaria.

L'efficienza relativa delle etichette complementari. Si definisce l'efficienza relativa di una etichetta complementare rispetto a una ordinaria come il rapporto tra l'informazione di Fisher che ciascuna etichetta porta sull'accuratezza vera A :

$$I_c(A, K) / I_o(A) = \text{Var}(\hat{A}_{\text{ord}}) / \text{Var}(\hat{A}_{\text{comp}}) |_{\{n_c=n_o=1\}} = A \cdot (K-1) / (A + K - 2)$$

dove si è usato il fatto che, per stimatori non distorti, la quantità di informazione di Fisher è inversamente proporzionale alla varianza (quando la varianza raggiunge il bound di Cramér-Rao, che non è garantito qui ma si verifica per stimatori lineari di Bernoulli). La formula mostra che l'efficienza relativa è:

$$\epsilon(A, K) = A(K-1) / (A + K - 2)$$

che ha le seguenti proprietà: per $K = 2$ si ha $\epsilon = A$, ossia le etichette complementari sono A volte meno efficienti delle ordinarie — per $A = 0.5$ (sistema casuale), $\epsilon = 0.5$; per $K \rightarrow \infty$ si ha $\epsilon \rightarrow A / (1 + A/(K-1)) \rightarrow A$

(l'efficienza saturà ad A per K grande); per $A \rightarrow 1$ si ha $\epsilon \rightarrow (K-1)/(K-1) = 1$, ossia quando il sistema è quasi perfetto le etichette complementari sono tanto efficienti quanto le ordinarie.

La derivata di ϵ rispetto a K è:

$$\partial\epsilon/\partial K = A(A + K - 2 - (K-1)) / (A + K - 2)^2 = A(A - 1) / (A + K - 2)^2$$

che è sempre negativa per $A < 1$. Ciò significa che l'efficienza relativa delle etichette complementari *decresce monotonicamente* con il numero di opzioni: aggiungere più opzioni al benchmark rende ogni singola etichetta complementare meno informativa. Intuitivamente, con K grande l'informazione che "l'opzione j non è la risposta giusta" è quasi irrilevante perché ci sono K-1 altre opzioni, e la probabilità che il sistema eviti accidentalmente l'opzione complementare è alta.

Il costo totale e il punto di ottimo. Sia c_o il costo di raccolta di una etichetta ordinaria e $c_c = c_r \cdot c_o$ il costo di raccolta di una etichetta complementare, con $c_r < 1$ (le etichette complementari sono più economiche). Per raggiungere una varianza target $Var^* = A(1-A)/n_{target}$ con budget totale $B = n_o \cdot c_o + n_c \cdot c_c$, il numero ottimale di etichette di ciascun tipo si ottiene risolvendo il problema di ottimizzazione vincolata:

$$\text{minimizzare } Var(\hat{A}_{mix}) = [A(1-A)/n_o + (A+K-2)(1-A)/n_c] / [A(1-A)/n_o + (A+K-2)(1-A)/n_c]$$

$$\text{soggetto a } n_o \cdot c_o + n_c \cdot c_c = B$$

La soluzione di questo problema di ottimo vincolato — attraverso il moltiplicatore di Lagrange — dà il rapporto ottimale:

$$n_c^* / n_o^* = \sqrt{[(A+K-2) / A]} / \sqrt{c_r} \cdot \sqrt{(1/c_r)} = \sqrt{[(A+K-2) / (A \cdot c_r^2)]}$$

e la varianza minima ottenibile con budget B è:

$$Var^*_{min}(B, K, A, c_r) = [A(1-A) \cdot (A+K-2)(1-A)] / [B/c_o] \cdot [\sqrt{A} + \sqrt{((A+K-2) \cdot c_r)}]^2$$

Derivando rispetto a K e ponendo la derivata uguale a zero, il numero ottimo di opzioni K^* che minimizza la varianza per budget fisso è:

$$K^* = \text{argmin}_K Var^*_{min}(B, K, A, c_r)$$

La condizione del primo ordine si semplifica a:

$$\partial/\partial K [\sqrt{(A+K-2)} \cdot (\sqrt{A} + \sqrt{((A+K-2) \cdot c_r)})^2] = 0$$

che, sviluppata, dà:

$$\sqrt{A} \cdot (A + K - 2)^{-1/2} / 2 + \sqrt{c_r} \cdot (3/2)(A + K - 2)^{1/2} \cdot \dots = 0$$

La soluzione analitica chiusa non è elementare per K generale, ma numericamente si trova che K^* cresce con c_r (se il costo relativo delle etichette complementari è alto, conviene avere poche opzioni) e decresce con A (se il sistema è molto accurato, conviene avere molte opzioni, poiché l'efficienza relativa $\epsilon \rightarrow 1$ per $A \rightarrow 1$ e K grande). Per valori tipici $A \in [0.5, 0.9]$ e $c_r \in [0.1, 0.5]$, il numero ottimale è $K^* \in [4, 8]$ — un range che corrisponde ai formati a scelta multipla standard della maggior parte dei benchmark di valutazione per LLM (MMLU, ARC, HellaSwag), suggerendo che il formato 4-opzioni non è soltanto una convenzione storica ma rispecchia un ottimo informativo approssimativo sotto ipotesi ragionevoli sul costo relativo dell'expertise.

16.7.1 Lo stimatore stratificato per fattori JEPa: riduzione della varianza per condizionamento sulla struttura causale latente

La connessione tra il framework di Yin et al. e il paradigma JEPa delineata nella sezione 16.7 identifica una struttura produttiva: la partizione dell'expertise degli annotatori rispecchia la fattorizzazione dello spazio latente JEPa. Questa corrispondenza non è soltanto metaforica — può essere sfruttata per costruire uno stimatore dell'accuratezza del sistema che sia strutturalmente più efficiente dello stimatore complementare standard, usando la struttura causale dello spazio JEPa come variabile di stratificazione.

Setup formale. Sia $f: \mathcal{X} \rightarrow \{1, \dots, K\}$ un sistema di classificazione a K opzioni basato su JEPa, con $z_AGG = g(E(x))$ la rappresentazione aggregata nello spazio latente. Il causal probing del Capitolo 12 ha identificato che la coordinata $z^{\{j\}}$ — il j -esimo fattore latente dello spazio JEPa — è la variabile più predittiva della risposta corretta Y attraverso il test di probing interventale su un dataset di validazione sperimentale. Sia $\varrho^2 = R^2_causal(z^{\{j\}}, Y)$ l' R^2 del probing causale di $z^{\{j\}}$ sulla variabile target Y , e sia $\varrho^2_perf = R^2_causal(z^{\{j\}}, \mathbb{1}[f(X) = Y])$ l' R^2 del probing causale di $z^{\{j\}}$ sulla performance binaria del sistema (corretta/errata). Quest'ultima quantità misura in che misura il fattore causale j è predittivo del *comportamento* del sistema, non soltanto del target vero.

Costruzione dello stimatore stratificato. Si partiziona il dataset di valutazione $\mathcal{D}_{eval} = \{x_i, \hat{y}_i, \bar{Y}_i\}_{i=1}^N$ in m strati $S_1, \dots, S_m$ in base al valore di $z^{\{j\}}(x_i)$:

$$S_l = \{x_i \in \mathcal{D}_{eval} : z^{\{j\}}(x_i) \in [v_l, v_{l+1})\}$$

dove $v_0 < v_1 < \dots < v_m$ sono i quantili di $z^{\{j\}}$ (ossia ogni strato contiene approssimativamente n/m campioni). All'interno di ogni strato S_l , si calcola la stima dell'accuratezza usando lo stimatore complementare intra-strato:

$$\hat{A}_l = (K-1) \cdot (1/n_l) \sum_{i \in S_l} W_i - (K-2) = (K-1) \cdot \hat{q}_l - (K-2)$$

dove $W_i = \mathbb{1}[f(x_i) \neq \tilde{Y}_i]$ e $n_l = |S_l|$ è la dimensione del l -esimo strato. Lo *stimatore stratificato per fattori JEP* è la media pesata degli stimatori intra-strato:

$$\hat{A}_{\text{JEP-strat}} = \sum_{l=1}^m (n_l / N) \cdot \hat{A}_l$$

Proprietà statistiche: non-distorsione e riduzione della varianza. Lo stimatore $\hat{A}_{\text{JEP-strat}}$ è non distorto: $\mathbb{E}[\hat{A}_{\text{JEP-strat}}] = \sum_l (n_l/N) \cdot A_l = A$, dove $A_l = P(f(X) = Y \mid z^{\{j\}}(X) \in [v_l, v_{l+1}))$ è l'accuratezza condizionale nello strato l . La non distorsione vale perché $\sum_l (n_l/N) \cdot A_l = P(f(X) = Y) = A$ per la legge della probabilità totale.

La varianza dello stimatore stratificato si decompone per la legge della varianza totale. Chiamando n_l la dimensione realizzata (aleatoria) dello strato l :

$$\text{Var}(\hat{A}_{\text{JEP-strat}}) = \sum_l (n_l/N)^2 \cdot \text{Var}(\hat{A}_l \mid n_l) = \sum_l (1/N) \cdot (A_l + K - 2)(1 - A_l) / (K-1)^2$$

dove si è usato il fatto che $\text{Var}(\hat{A}_l) = (A_l + K - 2)(1 - A_l) / n_l$ (la formula della sezione 16.3 applicata a ciascuno strato con accuratezza A_l).

Per confronto, la varianza dello stimatore non stratificato è:

$$\text{Var}(\hat{A}_{\text{comp}}) = (A + K - 2)(1 - A) / N$$

La riduzione della varianza ottenuta dalla stratificazione è:

$$\Delta_{\text{Var}} = \text{Var}(\hat{A}_{\text{comp}}) - \text{Var}(\hat{A}_{\text{JEP-strat}}) = (A + K - 2)(1 - A)/N - (1/N) \sum_l (n_l/N)(A_l + K - 2)(1 - A_l)$$

Usando la convessità della funzione $f(a) = (a + K - 2)(1 - a)$ (che è concava in a per $K \geq 2$), e la disuguaglianza di Jensen in direzione opposta:

$$\sum_l (n_l/N) f(A_l) \leq f(\sum_l (n_l/N) \cdot A_l) = f(A)$$

la disuguaglianza vale nella direzione giusta per garantire $\Delta_{\text{Var}} \geq 0$: stratificare non può mai *aumentare* la varianza. La varianza si riduce di più quando la funzione $f(A_l)$ varia molto tra strati — ossia quando la stratificazione espone grande eterogeneità nell'accuratezza tra strati.

Il bound di riduzione della varianza attraverso il probing R^2 . La quantità di riduzione della varianza è legata alla varianza tra strati dell'accuratezza:

$$\Delta_Var = (1/N)(1-A) \cdot Var_{{strati}}(A_l) + \text{termini di ordine superiore}$$

dove $Var_{{strati}}(A_l) = \sum_l (n_l/N)(A_l - A)^2$ è la varianza tra strati dell'accuratezza. Questa grandezza è direttamente collegata alla forza della correlazione tra il fattore j e la performance del sistema:

$$Var_{{strati}}(A_l) \leq A(1-A) \cdot \varrho^2_perf$$

dove $\varrho^2_perf = R^2_causal(z^{\{j\}}, \mathbb{1}[f(X) = Y])$ è l' R^2 del probing causale del fattore j sulla performance. La disuguaglianza segue dalla decomposizione della varianza di una variabile binaria attraverso un predittore continuo. Combinando le due formule:

$$\Delta_Var \leq A(1-A)^2 \cdot \varrho^2_perf / N$$

La riduzione della varianza è quindi proporzionale al quadrato del probing R^2 della performance: fattori che sono fortemente predittivi del comportamento del sistema (ϱ^2_perf alto) permettono stratificazioni più efficaci, con riduzioni di varianza più pronunciate. Questo risultato connette il probing causale — strumento di analisi interno al sistema JEPA — alla progettazione ottimale del protocollo di valutazione esterna: un buon probing diagnostico delle rappresentazioni JEPA può guidare direttamente la scelta delle variabili di stratificazione per il benchmark di valutazione.

La connessione con il teorema di Rao-Blackwell. Lo stimatore stratificato $\hat{A}_JEPA-strat$ è costruttivamente simile all'applicazione del Teorema di Rao-Blackwell: si condiziona lo stimatore grezzo $\hat{A}_comp$ sulla statistica sufficiente (approssimata) $Z^{\{j\}}$ per l'accuratezza del sistema, riducendo la varianza. La connessione non è esatta — il Teorema di Rao-Blackwell richiede che la variabile di condizionamento sia una statistica sufficiente nel senso di Fisher per il parametro A , mentre $Z^{\{j\}}$ ne è soltanto una proxy approssimata — ma la struttura del miglioramento è analoga: condizionare su una variabile correlata con la performance rimuove la varianza tra strati, lasciando soltanto la varianza residua intra-strato.

La differenza fondamentale rispetto all'applicazione standard di Rao-Blackwell è che $Z^{\{j\}}$ è una variabile *interna* al sistema — prodotta dall'encoder JEPA — non una variabile esterna al processo di raccolta dati. Il miglioramento dell'estimatore sfrutta quindi la struttura causale delle rappresentazioni del sistema per rendere più efficiente la sua stessa valutazione: un caso in cui la teoria della rappresentazione causale e la statistica della valutazione si coniugano in modo diretto e operativo.

Applicazione al benchmarking di sistemi visivi JEPA con annotatori specializzati. Lo stimatore stratificato per fattori JEPA si applica naturalmente al benchmarking di V-JEPA 2.1 in task di segmentazione semantica di immagini mediche. In questo contesto, il fattore latente più predittivo della performance del sistema potrebbe essere $z^{\{j\}} =$ la componente di z_AGG che

corrisponde alla struttura spaziale locale (identificata attraverso il probing con l'obiettivo di stima di profondità). La stratificazione su $z^{\{j\}}$ partiziona le immagini in strati di complessità strutturale crescente — immagini con strutture semplici nel primo strato, immagini con strutture complesse nell'ultimo — e permette di stimare separatamente l'accuratezza in ciascuna fascia di complessità.

Gli annotatori (specialisti medici) vengono assegnati alle immagini di ciascuno strato in modo stratificato, e le loro etichette complementari ("questo pixel non appartiene a una lesione di tipo X") vengono aggregate con lo stimatore $\hat{A}_l$ all'interno dello strato. L'aggregazione degli stimatori intra-strato produce $\hat{A}_{JEPA-strat}$ con varianza ridotta rispetto allo stimatore non stratificato, in misura proporzionale all'eterogeneità dell'accuratezza tra strati — che è tipicamente elevata nei task di segmentazione medica, dove la difficoltà delle immagini varia enormemente a seconda del tipo di patologia e della chiarezza delle strutture anatomiche.

Il risultato pratico è un protocollo di valutazione più efficiente dal punto di vista statistico: per raggiungere la stessa precisione dello stimatore complementare standard, lo stimatore stratificato richiede meno etichette complementari. Poiché ciascuna etichetta richiede il tempo e l'expertise di uno specialista medico, la riduzione del numero di etichette necessarie si traduce direttamente in una riduzione del costo di valutazione — che è spesso il vincolo dominante nella valutazione di sistemi AI in domini specialistici ad alta barriera di competenza.

16.4 Gli stimatori misti: combinazione ottima di etichette ordinarie e complementari

Nel regime pratico, si dispone tipicamente sia di alcune etichette ordinarie (costose, richiedono la verifica completa) sia di molte etichette complementari (economiche, richiedono soltanto che l'esperto risponda «non è nel mio dominio»). La combinazione ottimale di questi due tipi di etichette è il problema degli *stimatori misti* formalizzati da Yin et al.

Siano $\hat{A}_{ord} = S_o/n_o$ e $\hat{A}_{comp} = (K-1) \cdot S_c/n_c - (K-2)$ i due stimatori, entrambi non distorti e indipendenti (calcolati su set disgiunti). Lo stimatore misto lineare è:

$$\hat{A}_{mix} = w \cdot \hat{A}_{ord} + (1-w) \cdot \hat{A}_{comp}$$

che è anch'esso non distorto per qualsiasi $w \in [0,1]$. Minimizzando la varianza rispetto a w si ottiene il peso ottimo:

$$w^* = \text{Var}(\hat{A}_{comp}) / (\text{Var}(\hat{A}_{ord}) + \text{Var}(\hat{A}_{comp}))$$

ossia lo stimatore con peso proporzionale inverso alla varianza — il classico *Inverse Variance Weighting* (IVW). Il peso ottimo ha la forma chiusa:

$$w^* = ((A+K-2) \cdot n_o) / (A \cdot n_c + (A+K-2) \cdot n_o)$$

che dipende dall'incognita A . In pratica, si usa la stima pilota $\tilde{A} = \hat{A}_{\text{ord}}$ per calcolare il peso, ottenendo lo stimatore IVW plug-in. La varianza minima corrispondente è la media armonica delle varianze individuali:

$$\text{Var}_{\text{min}}(\hat{A}_{\text{mix}}) = \text{Var}(\hat{A}_{\text{ord}}) \cdot \text{Var}(\hat{A}_{\text{comp}}) / (\text{Var}(\hat{A}_{\text{ord}}) + \text{Var}(\hat{A}_{\text{comp}}))$$

Yin et al. propongono anche uno stimatore di massima verosimiglianza congiunta (ML) che usa sia $S_o \sim \text{Bin}(n_o, A)$ sia $S_c \sim \text{Bin}(n_c, q)$ per stimare A risolvendo l'equazione del punteggio. Il sistema di equazioni del punteggio si semplifica in un'equazione quadratica in A con coefficienti:

$$\alpha = N, \beta = (K-2)(T_o + T_c) + (K-3) \cdot S_o - S_c, \gamma = -(K-2) \cdot S_o$$

dove $T_o = n_o - S_o$, $T_c = n_c - S_c$, $N = n_o + n_c$. Lo stimatore ML è la radice unica in $[0,1]$ della forma quadratica:

$$\hat{A}_{\text{ML}} = (-\beta + \sqrt{\beta^2 - 4\alpha\gamma}) / (2\alpha)$$

La connessione formale tra ML e IVW è dimostrata nella Proposizione D del paper (Appendice D): i due stimatori sono equivalenti al primo ordine attraverso un'espansione Newton attorno a $A_0 = \hat{A}_{\text{ord}}$, il che conferma che l'IVW è un'approssimazione statistica efficiente dell'ML con calcolo più semplice.

16.5 Garanzie a campioni finiti: i bound di deviazione

La validità degli stimatori per campioni finiti — necessaria per applicazioni pratiche di scalable oversight dove i dataset di valutazione sono necessariamente di dimensione limitata — è garantita dai teoremi di Yin et al. attraverso bound di deviazione di tipo Hoeffding e Bernstein empirico.

Per lo stimatore complementare, il Teorema 2 fornisce con probabilità almeno $1 - \delta$:

$$|\hat{A}_{\text{comp}} - A| \leq (K-1) \cdot \min(\sqrt{\log(2/\delta) / (2 \cdot n_c)}, \sqrt{2 \cdot \hat{q}(1-\hat{q}) \cdot \log(4/\delta) / (n_c-1)} + 7 \cdot \log(4/\delta) / (3 \cdot (n_c-1)))$$

Il primo termine del min è il bound di Hoeffding — semplice e invariante alla varianza ma conservativo; il secondo è il bound di Bernstein empirico — adattivo alla varianza empirica $\hat{q}(1-\hat{q})$ e tipicamente più stretto quando $\hat{q}$ è vicino a 0 o 1.

Per lo stimatore misto con peso fisso w , il Teorema 4 fornisce un bound di tipo Bernstein che sfrutta entrambe le componenti di varianza:

$$|\hat{A}_{\text{mix}} - A| \leq \sqrt{(2v \cdot \log(2/\delta))} + c \cdot \log(2/\delta)$$

dove $v = w^2 \cdot A(1-A)/n_o + (1-w)^2 \cdot (K-1)^2 \cdot q(1-q)/n_c$ è la varianza totale del mix, e $c = \max(w/n_o, (1-w)(K-1)/n_c)$ è il termine di range. Questo bound si restringe come $O(1/\sqrt{n_o}) + O(1/\sqrt{n_c})$ all'aumentare di entrambi i campioni.

La dipendenza del bound dalla varianza v ha un'implicazione pratica importante: per il peso IVW ottimo w^* , la varianza v è già minimizzata, il che significa che il bound di Bernstein per $\hat{A}_{\text{mix}}$ con $w = \hat{w}_{\text{IVW}}$ è automaticamente il più stretto ottenibile nella classe degli stimatori lineari.

16.6 Validazione empirica: LLM benchmark senza ground truth

La validazione empirica del framework di Yin et al. copre tre setting distinti che illustrano le diverse modalità di applicazione.

Il primo setting è la validazione statistica su quattro benchmark LLM dove il ground truth è disponibile: MMLU-Pro (9795 item a 10 opzioni), MedQA-USMLE (a 4 opzioni), GPQA — il Graduate-Level Google-Proof Q&A benchmark di Rein et al. — (a 4 opzioni, 120 campioni), e MATH-MC (11751 item a 5 opzioni). Le etichette complementari sono costruite sinteticamente dal ground truth seguendo il protocollo: per ogni item, si seleziona uniformemente a caso uno degli indici non-corretti come etichetta complementare. Il modello valutato è gpt-5-nano (con gpt-4.1-nano per MATH-MC per evitare effetti di soffitto).

I risultati in Tabella 1 del paper mostrano che l'IVW e l'ML raggiungono la più bassa deviazione standard within-run tra tutti gli stimatori — indicatore diretto della qualità dello stimatore nella sua funzione primaria di riduzione della varianza. Per MMLU-Pro, l'IVW raggiunge $77.97\% \pm 1.58$ (deviazione standard across-seed) con deviazione within-run ± 1.79 , contro $77.00\% \pm 12.49$ (± 7.95) per lo stimatore complementare con $n_c = n_o$. La deviazione standard nell'accuratezza sull'intero dataset (Ord-Eval) è 77.97% , indicando che IVW con sole etichette complementari più poche etichette ordinarie replica il risultato della valutazione completa con incertezza comparabile.

Il secondo setting è la valutazione su task reali dove il ground truth è genuinamente inaccessibile per singoli esperti. EDINET-Bench (Sugiura et al. 2026) è un dataset di classificazione di report finanziari giapponesi in 16 settori industriali, dove un esperto in un settore può ragionevolmente escludere la propria specializzazione ma non identificare quella corretta tra le 15 rimanenti. Medical Abstracts (Schopf et al. 2023) è un dataset di classificazione di abstract medici per categoria di malattia, dove un clinico specialista può escludere la propria categoria ma raramente identificare con sicurezza quella corretta in una tassonomia a K categorie. I risultati in Tabella 2 mostrano che IVW e ML mantengono la deviazione standard più bassa (lowest variance across 3 seeds),

con IVW che produce la stima più vicina al riferimento ordinario su EDINET Extended e Medical Abstracts.

Il terzo setting — *agentic training with complementary feedback* — è il più innovativo rispetto allo stato dell'arte del *scalable oversight*: le etichette complementari vengono usate non soltanto per valutare il sistema ma come segnale di addestramento per ottimizzare workflow agentici automatici. ADAS (Automated Design of Agentic Systems, Hu et al. 2025) e AFlow (Zhang et al. 2025) sono pipeline di ricerca architetturale per sistemi agentici che usano un'accuratezza di validazione come funzione di fitness. Sostituendo l'accuratezza ordinaria con la accuratezza complementare (e la sua versione trasformata $\hat{A}_{comp}$), entrambe le pipeline producono sistemi con performance superiori ai baseline manualmente progettati su GPQA, Math-MC e Medical Abstracts, come mostrato in Figura 3.

16.7 Connessione con il paradigma JEPA: la partizione dell'expertise come inversione della factorizzazione latente

La connessione tra il framework di Yin et al. e il paradigma JEPA non è immediata — i due contributi vengono da tradizioni distinte — ma diventa produttiva quando si interpreta la *partizione dell'expertise* attraverso la lente della factorizzazione dello spazio latente.

In un sistema JEPA addestrato su un dominio complesso (per esempio la diagnosi medica da immagini), lo spazio latente z_{AGG} è strutturato in modo che diverse direzioni corrispondano a diverse proprietà dell'input — alcune legate all'anatomia (struttura ossea, pattern vascolare), altre alla patologia (bordi di lesioni, asimmetrie), altre all'artefatto di acquisizione (rumore MRI, contrasto). Un esperto cardiologico è, dal punto di vista dello spazio latente JEPA, qualcuno che ha alta affidabilità nell'identificare le direzioni latenti legate alla cardiologia e bassa affidabilità sulle direzioni latenti delle altre specializzazioni.

L'etichetta complementare fornita dal cardiologo — «questo non è un problema cardiovascolare» — può essere interpretata come un'affermazione sull'orientamento del vettore delle previsioni del sistema rispetto alle direzioni latenti cardiologiche: il vettore di previsione del sistema non è allineato con la sottospazio cardiologico dello spazio latente. Questa informazione è indiretta rispetto all'etichetta ordinaria (che direbbe esattamente a quale sottospazio il vettore di previsione è allineato), ma è comunque sistematicamente informativa grazie all'ipotesi di uniformità della distribuzione delle etichette complementari.

In questo quadro, il framework di Yin et al. diventa un meccanismo di *scalable oversight partizionato rispetto alla struttura dello spazio JEPA*: ciascun supervisore umano fornisce informazione sul sottospazio che non è responsabile delle previsioni, e la aggregazione di queste informazioni complementari permette di

stimare l'accuratezza del sistema senza che nessun singolo supervisore debba comprendere l'intera struttura del task.

Questa connessione suggerisce un'estensione naturale del framework di Yin et al. al dominio visivo: per un sistema di classificazione o segmentazione basato su V-JEPA 2.1, gli esperti umani potrebbero fornire etichette complementari sulle feature dense — «questo patch non corrisponde alla superficie di un tavolo» — che, aggregate attraverso lo stimatore misto, permetterebbero di valutare la qualità delle rappresentazioni dense del sistema senza che nessun singolo annotatore debba comprendere l'intera scena 3D.

16.8 Limiti del framework e punti di contatto con le tensioni del paradigma JEPA

Il framework di Yin et al. ha limiti strutturali che è importante riconoscere esplicitamente, anche nella prospettiva della presente monografia.

Il primo è il *vincolo del formato a scelta multipla*. Il framework assume che il task sia riformulabile come classificazione a K opzioni, il che è ragionevole per molti task di valutazione (domande a scelta multipla, classificazione di documenti, diagnosi differenziale) ma non per task aperti (generazione di testo libero, pianificazione motoria, dimostrazione matematica). Gli autori riconoscono questa limitazione e propongono la riformulazione di alcuni task aperti in formato strutturato, ma il confine tra task riformulabili e task genuinamente aperti non è sempre tracciabile in modo principiato.

Il secondo è l'*assunzione di uniformità della distribuzione delle etichette complementari*. Il protocollo di raccolta dati garantisce l'uniformità per costruzione, ma richiede una struttura organizzativa specifica (K esperti distinti, uno per dominio) che non è sempre disponibile. Se gli annotatori sono scelti in modo non uniforme — per esempio se certi domini hanno più esperti disponibili e quindi vengono interrogati più frequentemente — la distribuzione delle etichette complementari non è più uniforme e lo stimatore $\hat{A}_{\text{comp}}$ acquista un bias sistematico.

Il terzo limite, rilevante in modo specifico per il paradigma AMI, riguarda il *regime in cui la conoscenza umana collassa completamente*. Il framework di Yin et al. scala il problema di supervisione dal «nessun singolo esperto può valutare l'output» a «un ensemble di esperti con competenze complementari può stimare l'accuratezza». Ma in task genuinamente sovraumani — dove nessun essere umano ha la competenza specialistica necessaria anche soltanto per escludere un'opzione con affidabilità — anche le etichette complementari sono inaffidabili. In questo regime estremo, né RLHF né il framework di Yin et al. offrono una soluzione principata: il problema diventa strutturalmente irrisolvibile senza supervisione da sistemi più forti (debootstrap recursivo, come nel framework di Burns et al.).

Questa limitazione fondamentale delimita il perimetro di applicabilità del scalable oversight partizionato. Il framework di Yin et al. non è una soluzione al problema del superallineamento in senso pieno — è piuttosto un'estensione dello spazio di task su cui la supervisione umana è ancora operativamente possibile, anche nella sua forma debole di esclusione complementare. Per i sistemi AMI del futuro prossimo — sistemi che operano a livelli di expertise paragonabili ai migliori esperti umani in domini singoli — questo spazio esteso è già sufficiente per garantire una forma di verifica. Per i sistemi a lungo termine che superano l'expertise umana in modi strutturali, l'estensione necessaria richiede strumenti concettualmente diversi.

16.9 La struttura del grafo causale del scalable oversight partizionato

Una lettura delle proposte di scalable oversight attraverso il formalismo del Capitolo 1 rivela che la questione del superallineamento è, al suo nucleo, un problema di *identificabilità causale in presenza di supervisori parziali*. Il DAG del scalable oversight partizionato può essere schematizzato come segue (nella notazione SCM del Capitolo 1):

Siano Z_{true} la variabile latente che rappresenta il «contenuto vero» di una risposta (nel caso di un sistema diagnostico, la diagnosi corretta), $Z_{\text{sys}} = f(X)$ la previsione del sistema, E_k la competenza dell'esperto k (una variabile che specifica il sottodominio di expertise), e L_k l'etichetta fornita dall'esperto k (etichetta ordinaria o complementare a seconda della risposta).

Il DAG causale è:

$Z_{\text{true}} \rightarrow L_k$ (l'etichetta dipende causalmente dal contenuto vero) $E_k \rightarrow L_k$ (l'etichetta dipende causalmente dalla competenza dell'esperto) $Z_{\text{sys}} \rightarrow W = \mathbb{1}[Z_{\text{sys}} \neq L_k]$ (la variabile di successo dipende causalmente dalla previsione e dall'etichetta)

In questo DAG, E_k funge da variabile di confondimento: la competenza dell'esperto influenza sia la qualità dell'etichetta (gli esperti più competenti nel dominio dell'opzione k producono etichette più affidabili) sia indirettamente l'utilità del segnale di supervisione. L'assunzione di uniformità della distribuzione delle etichette complementari è equivalente all'assunzione di *ignorabilità di E_k rispetto all'indice dell'etichetta complementare*: la scelta dell'indice complementare è indipendente dalla competenza dell'esperto selezionato, perché la selezione è randomizzata.

Questa connessione formalizza perché il protocollo di Yin et al. funziona: la randomizzazione della selezione dell'esperto «apre» il backdoor path da Z_{true} a W attraverso E_k , rendendo identificabile la relazione causale tra Z_{true} e Z_{sys} attraverso la sola osservazione di W . Senza randomizzazione — se certi esperti vengono consultati più frequentemente per certi tipi di item — la

distribuzione di $p(W \mid Z_{\text{sys}})$ sarebbe distorta dalla selezione non uniforme, e lo stimatore $\hat{A}_{\text{comp}}$ acquisterebbe un bias che il framework teorico non riesce a correggere.

16.10 Implicazioni per il programma di ricerca sull'allineamento dei sistemi JEPA

Il contributo di Yin et al. (2026) — pubblicato come paper accettato a ICLR 2026, con codice disponibile — offre uno strumento operativo per una delle sfide più concrete del scalable oversight: come raccogliere e usare segnali di supervisione in domini dove la competenza umana è frammentata. Le sue implicazioni per il programma di ricerca sull'allineamento dei sistemi JEPA sono le seguenti.

Sul piano della *valutazione*, il framework permette di costruire benchmark di valutazione per sistemi visivi fondazionali — come V-JEPA 2.1 — che non richiedono annotatori con competenze enciclopediche. Per un sistema di segmentazione semantica in un dominio specialistico (diagnostica per immagini, ispezione industriale, guida autonoma), esperti con competenze in sotto-domini specifici potrebbero fornire etichette complementari («questo pixel non appartiene a una lesione dermatologica») che, aggregate, permettono di stimare l'accuratezza del sistema in modo statisticamente valido.

Sul piano del *training agentivo*, la dimostrazione di Yin et al. che le etichette complementari possono essere usate come segnale di fitness per la ricerca architetturale automatica (ADAS, AFlow) apre una direzione per l'addestramento di sistemi JEPA in domini dove il ground truth è inaccessibile. Un sistema JEPA per la percezione robotica in ambienti industriali potrebbe essere ottimizzato con etichette complementari fornite da operatori con competenze settoriali, senza richiedere supervisor che abbiano familiarità con l'intero dominio operativo del robot.

Sul piano dei *limiti strutturali*, il framework conferma la traiettoria del superallineamento come la sfida irrisolvibile entro il paradigma della supervisione umana diretta. In ultima analisi, la scalable oversight partizionata estende il raggio operativo della supervisione umana, ma non lo rende illimitato. Il punto di rottura — dove nessun esperto umano può fornire nemmeno etichette complementari affidabili — è il confine del paradigma. Oltre quel confine, la ricerca sull'allineamento dovrà sviluppare strumenti concettualmente diversi: metodi di *auto-consistenza* che sfruttino la struttura interna del sistema per generare i propri segnali di supervisione, o metodi di *verifica formale* che traducano proprietà dei sistemi in affermazioni dimostrabili indipendentemente dall'expertise del valutatore umano. Il paradigma JEPA, con la sua enfasi sulla struttura causale esplicita degli spazi latenti e sulla verificabilità esterna attraverso probing e TCAV, offre le fondamenta architetturali più promettenti per entrambe le direzioni.

La Quarta Parte ha tracciato il territorio che separa la competenza tecnica dall'allineamento verificabile. Il probing causale, TCAV e gli Sparse Autoencoder hanno fornito gli strumenti per rendere ispezionabili gli spazi latenti dei sistemi JEPA: non nel senso di renderli trasparenti — il che richiederebbe una corrispondenza biunivoca tra le dimensioni dello spazio latente e i concetti umani, una condizione che la geometria degli spazi ad alta dimensionalità non garantisce spontaneamente — ma nel senso di renderli interrogabili rispetto a ipotesi specifiche sulla struttura causale della rappresentazione. Il framework RLHF, formalizzato come ottimizzazione vincolata nel Capitolo 15, ha mostrato come il feedback umano possa essere impiegato per ancorare il comportamento del sistema al giudizio normativo dei valutatori, e ne ha derivato le condizioni di compatibilità con il training objective JEPA. Il scalable oversight del Capitolo 16, infine, ha identificato il punto di rottura del paradigma: il confine oltre il quale nessuna partizione dell'expertise umana è sufficiente a supervisionare un sistema AMI, e dove la richiesta di allineamento diventa una domanda senza risposta nell'orizzonte degli strumenti oggi disponibili.

Tuttavia, attraverso tutta la Quarta Parte ha operato un'assunzione implicita che non è mai stata formalizzata: che i fattori latenti rispetto ai quali il probing causale opera, le direzioni lungo cui TCAV misura l'attivazione dei concetti, le componenti sparse che gli SAE decompongono, corrispondano effettivamente a variabili causalmente rilevanti nella struttura del mondo che il sistema percepisce e in cui agisce. Questa assunzione è il ponte tra l'interpretabilità e la causalità, e il suo fondamento formale non è mai stato derivato esplicitamente nella presente monografia: è stato invocato, nella Prima Parte, come motivazione del programma del Causal Representation Learning, e nuovamente invocato, nel Capitolo 12, come ipotesi di lavoro per il probing causale. Ricapitolare questa posizione non basta: le quattro parti precedenti hanno costruito un edificio le cui fondamenta causali restano implicite nelle assunzioni di ciascun capitolo piuttosto che esplicitate in un risultato unificante. Il Capitolo 17 è dedicato a colmare questa lacuna derivativa, mostrando che la struttura matematica di Var-JEPA non invoca la causalità come metafora o come aspirazione, ma la contiene come conseguenza derivabile della sua struttura generativa — e che questa conseguenza ha implicazioni architetturali verificabili che le quattro parti precedenti hanno presupposto senza dimostrare.

Capitolo 17 — La congiunzione formale: distribuzioni interventali e DAG generativo come programma di ricerca unificante

17.1 Il DAG implicito nelle architetture JEPA: dall'architettura alla struttura causale

Attraversare i sedici capitoli precedenti con occhio ricostruttivo rivela una tensione persistente che ne attraversa l'intera struttura argomentativa: il paradigma JEPA ha sistematicamente *invocato* la causalità come criterio di qualità delle rappresentazioni — le buone rappresentazioni catturano strutture causali, i World Models modellano dipendenze e non mere correlazioni, lo spazio latente ideale separa i fattori che variano per ragioni indipendenti — senza mai aver *derivato* formalmente come si costruisce un modello causale sui fattori latenti di una architettura concreta. Il Capitolo 17 è dedicato a colmare questa lacuna derivativa: l'obiettivo è mostrare che la struttura matematica di Var-JEPA contiene già, in forma latente, un grafo aciclico diretto (DAG) causale che può essere reso esplicito e, attraverso le recenti proposte architetturelle di LeWorldModel, seq-JEPA e V-JEPA 2.1, avvicinato alla struttura degli Structural Causal Models di Pearl in modo operativamente utile.

Il punto di partenza è la struttura dell'obiettivo JEPA nella sua formulazione canonica. Siano $o_{1:T} \in \mathcal{O}^T$ le osservazioni grezze lungo una traiettoria di lunghezza T , e siano $a_{1:T} \in \mathcal{A}^T$ le azioni associate. L'encoder $\text{enc}_\theta : \mathcal{O} \rightarrow \mathbb{R}^d$ mappa ogni osservazione in uno spazio latente d -dimensionale, producendo $z_t = \text{enc}_\theta(o_t)$. Il predittore $\text{pred}_\varphi : \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}^d$ stima il successore latente come $\hat{z}_{t+1} = \text{pred}_\varphi(z_t, a_t)$. L'obiettivo di addestramento nella formulazione più semplice — quella che sarà precisata nella sottosezione 17.3 con riferimento a LeWorldModel — è la somma di un termine di previsione e un termine di regolarizzazione:

$$\mathcal{L}_{\text{JEPA}} \triangleq \mathbb{E}_t [\|\hat{z}_{t+1} - z_{t+1}\|_2^2] + \lambda \cdot \text{Reg}(Z)$$

dove $Z \in \mathbb{R}^{N \times B \times d}$ è il tensore degli embedding raccolti sull'intera sequenza e sul batch, N è la lunghezza della storia, B la dimensione del batch, e $\text{Reg}(Z)$ è il termine di regolarizzazione anti-collasso, la cui forma specifica sarà oggetto della sottosezione 17.3.

La struttura causale implicita in questo obiettivo emerge quando si chiedono le relazioni di dipendenza tra le variabili. Il DAG implicito di JEPA ha la seguente forma: siano S_t le variabili latenti che rappresentano lo stato del mondo al tempo t (inosservabili nella loro completezza), o_t le osservazioni prodotte da S_t attraverso una funzione di generazione $p(o_t | S_t)$, a_t le azioni che il sistema compie o osserva (trattate come esogene nel regime offline reward-free), e $z_t = \text{enc}_\theta(o_t)$ le rappresentazioni latenti prodotte dall'encoder.

Le relazioni di dipendenza sono le seguenti, nell'ordine causale: $S_t \rightarrow o_t$ (le osservazioni sono causate dallo stato del mondo); $o_t \rightarrow z_t$ (le rappresentazioni sono funzioni deterministiche delle osservazioni); $S_t \times a_t \rightarrow S_{t+1}$ (la dinamica del mondo è governata dallo stato corrente e

dall'azione); $z_t \times a_t \rightarrow \hat{z}_{t+1}$ (il predittore produce lo stimatore del successore latente). L'obiettivo JEPA forza la corrispondenza $\hat{z}_{t+1} \approx z_{t+1} = \text{enc}_\theta(o_{t+1})$, e quindi forza la corrispondenza tra la dinamica appresa nel latente e la dinamica reale nel mondo.

Questo DAG ha tre proprietà strutturali che lo rendono un oggetto causale, non meramente statistico. La prima è l'*asimmetria temporale*: le frecce di dipendenza vanno dal tempo t al tempo $t+1$, mai nella direzione opposta, il che riflette l'asimmetria causale del tempo fisico. La seconda è il *blocco markoviano*: condizionato su z_t e a_t , il successore $\hat{z}_{t+1}$ è statisticamente indipendente da z_{t-k} per ogni $k \geq 1$ — la struttura di markov di primo ordine è una proprietà caratteristica dei modelli causali dove lo stato corrente *maschera* il passato rispetto al futuro. La terza è la *separazione tra meccanismo di osservazione e meccanismo di transizione*: enc_θ e pred_φ sono componenti distinte con parametri separati, il che riflette la separazione tra il meccanismo che genera le osservazioni dal mondo e il meccanismo che governa la transizione del mondo.

Queste tre proprietà non sono accidentali rispetto alla struttura degli SCM: sono esattamente le proprietà che caratterizzano un modello di Markov causale nel senso di Pearl e Peters. La domanda è: quanto strettamente questa struttura si avvicina a un vero SCM, e in quali condizioni le distribuzioni interventali computate tramite il do-calcolo sul DAG JEPA corrispondono alle distribuzioni interventali sul mondo reale?

17.2 La distribuzione interventale come oggetto computabile: il do-calcolo nell'obiettivo JEPA

Il do-calcolo di Pearl permette di computare la distribuzione di una variabile Y sotto l'intervento $\text{do}(X = x)$ — ossia quando si impone il valore x alla variabile X tagliando le frecce causali entranti in X nel DAG — a partire dalla distribuzione osservazionale, a condizione che il DAG sia corretto e che il backdoor criterion (o equivalentemente il frontdoor criterion, o la c-component factorization) sia soddisfatto.

Nel contesto JEPA, l'intervento naturale da considerare è $\text{do}(a_t = a)$: si impone un'azione specifica a al tempo t , indipendentemente dalla distribuzione delle azioni nel dataset di addestramento $\pi(a_t \mid o_t)$. La distribuzione interventale di interesse è:

$$p(z_{t+1} \mid z_t, \text{do}(a_t = a)) \triangleq \int p(z_{t+1} \mid z_t, a_t = a, S_t) \cdot p(S_t \mid z_t) dS_t$$

Nel regime in cui il DAG JEPA è corretto — ossia quando l'encoder enc_θ cattura con sufficiente fedeltà le variabili latenti rilevanti dello stato S_t , in modo che la distribuzione $p(S_t \mid z_t)$ sia concentrata attorno ai valori di S_t consistenti con z_t — questa distribuzione interventale si semplifica in:

$$p(z_{t+1} \mid z_t, \text{do}(a_t = a)) \approx p(z_{t+1} \mid z_t, a_t = a)$$

ossia la distribuzione condizionale osservazionale. Il motivo per cui questa semplificazione è valida è che, nel setting offline reward-free con azioni esogene, le azioni a_t sono già trattate come variabili esogene — non c'è un meccanismo di selezione delle azioni che dipende dallo stato S_t in modo non osservato, o più precisamente, anche se esiste tale meccanismo, la variabile z_t già «schermizza» la dipendenza tra a_t e S_t attraverso il percorso $a_t \leftarrow \pi(a_t \mid o_t) \leftarrow o_t \rightarrow z_t$.

Questa semplificazione ha una conseguenza cruciale per la pratica: il predittore $\text{pred}_\varphi(z_t, a_t)$ apprende direttamente la distribuzione d'intervento, non soltanto la distribuzione condizionale osservazionale, a condizione che le azioni nel dataset di addestramento siano sufficientemente variegate da coprire lo spazio delle azioni rilevanti. Questa è la ragione per cui i sistemi JEPA addestrati in regime offline su dati eterogenei generalizzano al controllo in modo migliore di quanto la teoria del reinforcement learning osservazionale prevedrebbe: stanno apprendendo approssimazioni delle distribuzioni interventali, non mere correlazioni.

La formalizzazione del problema di generalizzazione può essere resa precisa attraverso il framework della *structural stability* di Peters, Mooij e Schölkopf. Sia $\mathcal{M}_{\text{JEPA}}$ il modello strutturale implicito, ossia il sistema di equazioni strutturali $\{z_{t+1} := \text{pred}_\varphi(z_t, a_t) + \epsilon_{t+1} : \epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2 I)\}$ dove l'errore residuo $\epsilon_{t+1} = z_{t+1} - \hat{z}_{t+1}$ cattura le deviazioni tra la previsione e la realtà. Il modello è *strutturalmente stabile* rispetto a interventi su a_t nella misura in cui la funzione $\text{pred}_\varphi(z_t, \cdot)$ è la stessa per tutte le azioni $a \in \mathcal{A}$, ossia nella misura in cui il predittore non ha memorizzato la correlazione specifica tra a_t e z_{t+1} nel dataset di addestramento, ma ha appreso la funzione causale che governa la transizione.

Il collegamento con la regolarizzazione anti-collasso diventa visibile attraverso questa lente. Se il predittore collassa (ossia se l'encoder mappa tutte le osservazioni allo stesso punto $z = c \in \mathbb{R}^d$), allora il modello strutturale diventa $\{z_{t+1} := c + \epsilon_{t+1}\}$, che è strutturalmente stabile in modo banale (la funzione pred_φ non dipende da a_t , poiché $z_{t+1} = c$ comunque) ma è informativamente vuoto. La regolarizzazione anti-collasso è quindi il meccanismo che forza il modello strutturale a essere *non-banalmente stabile*, ossia stabile rispetto a interventi in modo tale che la funzione $\text{pred}_\varphi(z_t, \cdot)$ sia effettivamente dipendente da a_t in modo causalmente significativo.

17.3 SIGReg come realizzazione computazionale del Principio di Meccanismi Causali Indipendenti

Il Principio di Meccanismi Causali Indipendenti (ICM) — enunciato da Peters et al. (2017) e da Schölkopf et al. (2021) come uno dei principi fondativi del Causal Representation Learning — afferma che i meccanismi causali che governano le diverse variabili in un sistema sono autonomi: la conoscenza del meccanismo che governa $P(X_i | Pa_i)$ non fornisce informazione sul meccanismo che governa $P(X_j | Pa_j)$ per $i \neq j$. In termini operativi, il principio implica che interventi su un meccanismo non modificano gli altri meccanismi — un'affermazione di modularità causale.

La realizzazione computazionale di questo principio nello spazio latente di un sistema JEPA richiede che i fattori latenti $z^{\{1\}}, z^{\{2\}}, \dots, z^{\{d\}}$ siano indipendenti in qualche senso preciso. La versione più forte è l'indipendenza marginale: $P(z^{\{i\}}) \perp P(z^{\{j\}})$ per $i \neq j$, ossia le distribuzioni marginali dei singoli fattori sono indipendenti. La distribuzione Gaussiana isotropa $\mathcal{N}(0, I_d)$ soddisfa esattamente questa condizione: sotto $P_Z = \mathcal{N}(0, I_d)$, si ha $z^{\{i\}} \perp z^{\{j\}}$ per $i \neq j$, poiché la matrice di covarianza è la matrice identità e le componenti sono gaussiane indipendenti.

Il Sketched-Isotropic-Gaussian Regularizer (SIGReg) di Balestriero e LeCun, introdotto in LeWorldModel come meccanismo anti-collasso, è la realizzazione computazionale di questa condizione. La sua definizione, richiamata per completezza, è:

$$\text{SIGReg}(Z) \triangleq (1/M) \sum_{m=1}^M T(Zu^{\{m\}})$$

dove $u^{\{m\}} \in S^{d-1}$ sono M direzioni casuali sull'ipersfera, $h^{\{m\}} = Zu^{\{m\}} \in \mathbb{R}^{N \cdot B}$ sono le proiezioni degli embedding lungo ciascuna direzione, e $T(\cdot)$ è la statistica test di Epps-Pulley:

$$T(h^{\{m\}}) = \int_{-\infty}^{+\infty} w(t) \cdot |\varphi_N(t; h^{\{m\}}) - \varphi_0(t)|^2 dt$$

con $\varphi_N(t; h) = (1/N) \sum_{n=1}^N e^{i t \cdot h_n}$ la funzione caratteristica empirica degli N campioni e $\varphi_0(t) = e^{-t^2/2}$ la funzione caratteristica della $\mathcal{N}(0,1)$. La funzione peso $w(t) = e^{-t^2/(2\lambda^2)}$ localizza il confronto nella frequenza, con λ che governa la scala di confronto.

La connessione con il principio ICM si stabilisce attraverso il teorema di Cramér-Wold: $\text{SIGReg}(Z) \rightarrow 0$ se e soltanto se $P_Z \rightarrow \mathcal{N}(0, I_d)$ in distribuzione. Poiché la Gaussiana isotropa ha fattori indipendenti, minimizzare $\text{SIGReg}(Z)$ forza la distribuzione degli embedding verso una distribuzione a fattori indipendenti — ossia realizza computazionalmente, in modo differenziabile e scalabile, la condizione di indipendenza marginale richiesta dal principio ICM.

È importante precisare il tipo di indipendenza realizzata. SIGReg non garantisce l'indipendenza *condizionale* dei fattori latenti dato qualsiasi valore di a_t — una condizione molto più forte, equivalente alla disentanglement nel senso di Eastwood e Williams (2018) — ma soltanto l'indipendenza *marginale*

dei fattori latenti integrati sulla distribuzione delle azioni. Questa forma più debole di indipendenza è, tuttavia, sufficiente per garantire le proprietà di generalizzazione interventale discusse nella sottosezione 17.2, poiché il backdoor criterion è soddisfatto quando le azioni a_t sono esogene rispetto ai fattori latenti marginalmente indipendenti.

Formalmente, sia la regola di aggiustamento backdoor di Pearl:

$$p(z_{t+1} \mid \text{do}(a_t = a)) = \sum \{z_t\} p(z_{t+1} \mid z_t, a_t = a) \cdot p(z_t)$$

che si semplifica (sotto l'ipotesi che z_t mascheri il percorso backdoor tra a_t e z_{t+1}) in:

$$p(z_{t+1} \mid \text{do}(a_t = a)) = \int \{z_t\} p(z_{t+1} \mid z_t, a_t = a) \cdot p_{\text{SIGReg}}(z_t) dz_t$$

dove $p_{\text{SIGReg}}(z_t)$ è la distribuzione marginale degli embedding imposta dalla regolarizzazione SIGReg, approssimata da $\mathcal{N}(0, I_d)$. La Gaussiana isotropa come distribuzione marginale degli stati latenti non è quindi una scelta cosmetica di convenienza computazionale: è la distribuzione che rende il calcolo delle distribuzioni interventistiche eseguibile in forma chiusa attraverso la regola di aggiustamento backdoor.

17.3.1 SIGReg come test non parametrico di normalità su proiezioni univariate

La costruzione matematica di SIGReg merita di essere sviluppata più esplicitamente, poiché la sua connessione con la teoria delle funzioni caratteristiche è il fondamento della sua ottimizzabilità attraverso backpropagation. Ricordando la definizione, la statistica di Epps-Pulley per un campione $h = (h_1, \dots, h_N)$ è:

$$T(h) = \int_{-\infty}^{+\infty} w(t) \cdot |\varphi_N(t;h) - \varphi_0(t)|^2 dt$$

dove $\varphi_N(t;h) = (1/N) \sum \exp(it h)$ è la funzione caratteristica empirica dei campioni e $\varphi_0(t) = \exp(-t^2/2)$ è la funzione caratteristica della $\mathcal{N}(0,1)$. Espandendo il quadrato del modulo:

$$|\varphi_N(t;h) - \varphi_0(t)|^2 = |\varphi_N|^2 - 2\text{Re}[\varphi_N \cdot \varphi_0^*] + |\varphi_0|^2$$

e integrando termine per termine con il peso $w(t) = \exp(-t^2/(2\lambda^2))$:

$$T(h) = (1/N^2) \sum_{m,n} K(h - h) - (2/N) \sum L(h) + C$$

dove $K(x) = \int w(t) \cdot \exp(itx) dt = \sqrt{2\pi} \cdot \lambda \cdot \exp(-x^2\lambda^2/2)$ è il kernel gaussiano con larghezza di banda λ (la trasformata di Fourier del peso), $L(x) = \int w(t) \cdot \text{Re}[\exp(itx)] \cdot \exp(-t^2/2) dt = \sqrt{\pi} \cdot (\lambda / (\sqrt{1/\lambda^2 + 1})) \cdot \exp(-x^2/(2(1+1/\lambda^2)))$ è un kernel misto, e $C = \int w(t) \cdot \exp(-t^2) dt = \sqrt{\pi/2} \cdot \lambda$ è una costante che non

dipende dai campioni. La forma kernel di $T(h)$ la rende differenziabile rispetto ai campioni h — e quindi rispetto ai parametri del modello che producono gli embedding — e calcolabile in $O(N^2)$ operazioni (o $O(N \log N)$ con approssimazioni FFT).

La media su M direzioni casuali produce $SIGReg(Z) = (1/M) \sum T(Zu^{(m)})$, che è anch'essa differenziabile e calcolabile in $O(M \cdot N^2)$ operazioni. La scelta di M è un iperparametro che bilancia precisione della stima (maggiore M) e costo computazionale (lineare in M): in LeWorldModel, Maes et al. usano $M = 512$ direzioni, che si sono rivelate sufficienti per stabilizzare il training su tutti i task considerati. La convergenza di $SIGReg(Z) \rightarrow 0$ è garantita dal teorema di Cramér-Wold — il quale afferma che una distribuzione multivariata è determinata dalle sue proiezioni univariate — e dall'analisi della potenza della statistica di Epps-Pulley come test non parametrico di normalità, che Epps e Pulley (1983) hanno dimostrato essere consistente: $T(h) \rightarrow 0$ se e soltanto se h è campionato da una $\mathcal{N}(0,1)$, con potenza asintotica unitaria per qualsiasi alternativa non gaussiana.

Questa derivazione ha una conseguenza diretta per la progettazione di sistemi JEPA causalmente informati: la scelta del regolarizzatore anti-collasso non è neutra rispetto alle proprietà causali del modello. Regolarizzatori che non impongono indipendenza marginale dei fattori — per esempio i regolarizzatori basati su VICReg (invarianza-varianza-covarianza) usati in PLDM, che impongono separatamente la varianza e la decorrelazione ma non la normalità — producono modelli che soddisfano le condizioni del principio ICM in modo più debole e quindi garantiscono proprietà di generalizzazione d'intervento più fragili.

Più precisamente: VICReg impone che la matrice di covarianza degli embedding abbia diagonale almeno 1 (il termine di varianza), che gli elementi fuori diagonale siano vicini a zero (il termine di covarianza), e che gli embedding di coppie positive siano simili (il termine di invarianza). Queste condizioni non implicano la normalità marginale: una distribuzione uniforme sull'ipersfera soddisfa la decorrelazione ma non è una Gaussiana. SIGReg, invece, impone direttamente la convergenza in distribuzione alla Gaussiana isotropa, che implica sia la decorrelazione sia la normalità delle proiezioni univariate — una condizione strettamente più forte di quella imposta da VICReg.

17.3.2 La forma esplicita del gradiente di SIGReg e la sua calcolabilità per backpropagation

La sezione 17.3.1 ha stabilito che la statistica di Epps-Pulley $T(h)$ per un campione univariato $h = (h_1, \dots, h_N)$ ammette la rappresentazione kernel:

$$T(h) = (1/N^2) \sum_{m,n} K(h_m - h_n) - (2/N) \sum_n L(h_n) + C$$

con $K(x) = \sqrt{2\pi} \cdot \lambda \cdot \exp(-x^2\lambda^2/2)$, $C = \sqrt{\pi/2} \cdot \lambda$, e il kernel misto $L(x)$ ottenuto come trasformata di Fourier del prodotto del peso w e della funzione caratteristica della Normale standard. È utile derivare la forma esplicita di $L(x)$ attraverso l'integrazione diretta, poiché questa forma è necessaria per l'implementazione differenziabile del regolarizzatore.

Il kernel misto $L(x)$ è definito come:

$$L(x) = \int_{-\infty}^{+\infty} w(t) \cdot \cos(tx) \cdot \exp(-t^2/2) dt = \int_{-\infty}^{+\infty} \exp(-t^2/(2\lambda^2)) \cdot \cos(tx) \cdot \exp(-t^2/2) dt$$

completando il quadrato nell'esponente: $-t^2/(2\lambda^2) - t^2/2 = -t^2(1 + \lambda^2)/(2\lambda^2)$, e usando l'integrale gaussiano standard con argomento complesso:

$$L(x) = \int_{-\infty}^{+\infty} \exp(-t^2(1 + \lambda^2)/(2\lambda^2)) \cdot \cos(tx) dt$$

Questo integrale è la parte reale della trasformata di Fourier di una Gaussiana di varianza $\sigma^2_{\text{eff}} = \lambda^2/(1 + \lambda^2)$, valutata in x :

$$L(x) = \sqrt{2\pi \cdot \lambda^2/(1 + \lambda^2)} \cdot \exp(-x^2 \cdot (1 + \lambda^2) / (2\lambda^2)) = \lambda \cdot \sqrt{2\pi/(1 + \lambda^2)} \cdot \exp(-x^2(1 + \lambda^2)/(2\lambda^2))$$

Questa forma ha la struttura di una Gaussiana scalata con varianza $1/(1 + \lambda^2)$ in x (o equivalentemente una Gaussiana di larghezza di banda $\lambda/\sqrt{1 + \lambda^2}$).

Con queste espressioni esplicite per K , L , e C , la statistica $T(h)$ è una funzione differenziabile di h . Il gradiente $\partial T/\partial h_n$ per un singolo campione h_n è:

$$\partial T/\partial h_n = (2/N^2) \sum_{j=1}^N \partial K(h_n - h_j)/\partial h_n - (2/N) \cdot \partial L(h_n)/\partial h_n$$

con:

$$\partial K(h_n - h_j)/\partial h_n = -\lambda^3 \cdot \sqrt{2\pi} \cdot (h_n - h_j) \cdot \exp(-(h_n - h_j)^2\lambda^2/2)$$

$$\partial L(h_n)/\partial h_n = -\lambda(1 + \lambda^2)/\lambda^2 \cdot \sqrt{2\pi/(1 + \lambda^2)} \cdot h_n \cdot \exp(-h_n^2(1 + \lambda^2)/(2\lambda^2)) = -\sqrt{2\pi} \cdot (1 + \lambda^2)^{1/2} \cdot \lambda^{-1} \cdot h_n \cdot \exp(-h_n^2(1 + \lambda^2)/(2\lambda^2))$$

Il gradiente $\partial T/\partial h_n$ è quindi la differenza di due somme: una somma di termini repulsivi (provenienti dal kernel K , che penalizza le coppie di campioni troppo vicine) e un termine attrattivo verso lo zero (proveniente da L , che incentiva i campioni a concentrarsi attorno alla media Gaussiana nulla). Questa struttura competitiva è la manifestazione a livello di gradiente del test di Epps-Pulley: la statistica T è nulla quando la distribuzione empirica è esattamente Gaussiana standard, ossia quando le forze repulsive e attrattive sono in equilibrio.

Il gradiente di SIGReg rispetto agli embedding. Il gradiente del regolarizzatore SIGReg rispetto all'embedding $z_n \in \mathbb{R}^d$ si ottiene attraverso la regola della catena. Ricordando che $\text{SIGReg}(Z) = (1/M) \sum_{m=1}^M T(Zu^{\{(m)\}})$ e che la proiezione scalare del campione n nella direzione m è $h_n^{\{(m)\}} = z_n^\top u^{\{(m)\}}$, si ha:

$$\frac{\partial \text{SIGReg}}{\partial z_n} = (1/M) \sum_{m=1}^M \frac{\partial T(Zu^{\{(m)\}})}{\partial h_n^{\{(m)\}}} u^{\{(m)\}} \cdot$$

dove $\partial T / \partial h_n^{\{(m)\}}$ è lo scalare derivato sopra, valutato sulla proiezione $Zu^{\{(m)\}}$. Questo gradiente è un vettore in $\mathbb{R}^d$ costruito come combinazione lineare delle M direzioni di proiezione $u^{\{(m)\}}$, pesate dai gradienti scalari corrispondenti. Il costo computazionale del calcolo è $O(M \cdot N \cdot d)$ per il forward pass (proiettare tutti gli N embedding nelle M direzioni e calcolare le N statistiche scalari) e $O(M \cdot N \cdot d)$ per il backward pass (propagare i gradienti scalari attraverso le proiezioni). Con $N = B$ (dimensione del batch, tipicamente 256–2048), $d = 256$ –512 (dimensione degli embedding), e $M = 512$ (numero di proiezioni casuali), il costo è comparabile a quello di un singolo layer Transformer di dimensione analoga — sufficientemente economico da essere usato come loss di regolarizzazione senza rallentamenti significativi del training.

Un punto tecnico rilevante per l'implementazione: le direzioni $u^{\{(m)\}}$ sono *campionate casualmente e poi congelate* per l'intera durata del training, non ricampionate ad ogni passo. Questa scelta è motivata da due ragioni. La prima è la consistenza del segnale: campionare nuove direzioni ad ogni passo introdurrebbe varianza aggiuntiva nel gradiente, mentre direzioni fisse producono un segnale di gradiente coerente attraverso gli aggiornamenti consecutivi. La seconda è l'equivalenza distributiva: per il teorema di Cramér-Wold, un numero sufficiente di direzioni fisse campionate uniformemente sull'ipersfera S^{d-1} è rappresentativo dell'intera distribuzione delle proiezioni univariate — non è necessario che le direzioni coprano *tutte* le possibili proiezioni, ma soltanto un sottoinsieme sufficientemente ricco ($M = 512$ è un valore ampiamente sufficiente per $d \leq 512$, come dimostrato negli esperimenti di ablation di LeWorldModel).

La connessione con il Maximum Mean Discrepancy e i kernel methods.

La forma kernel di $T(h)$ permette di riconoscere SIGReg come caso speciale del Maximum Mean Discrepancy (MMD) — la discrepanza statistica tra due distribuzioni misurata attraverso un kernel a caratteristica positiva — adattato al confronto tra la distribuzione empirica degli embedding e la distribuzione target $\mathcal{M}(0, I_d)$. Il MMD^2 tra due distribuzioni P e Q rispetto al kernel k è:

$$\text{MMD}^2(P, Q; k) = \mathbb{E}_{\{z, z' \sim P\}}[k(z, z')] - 2 \cdot \mathbb{E}_{\{z \sim P, \hat{z} \sim Q\}}[k(z, \hat{z})] + \mathbb{E}_{\{\hat{z}, \hat{z}' \sim Q\}}[k(\hat{z}, \hat{z}')] \cdot$$

Nell'approssimazione di Cramér-Wold, il kernel multivariato $k(z, z')$ viene scomposto in somme di kernel univariati sulle proiezioni: $k(z, z') \approx (1/M) \sum_m K(z^\top u^{\{(m)\}} - z'^\top u^{\{(m)\}})$, e i termini dell'aspettativa rispetto alla Gaussiana

isotropa $\hat{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ ammettono forma chiusa (essendo le proiezioni di una Gaussiana isotropa anch'esse Gaussiane univariate standard). La statistica $T(\mathbf{h})$ è esattamente questo MMD² unidimensionale sulla proiezione $\mathbf{h} = \mathbf{Z}\mathbf{u}$, e SIGReg è la media di M tali MMD² unidimensionali, connettendo la costruzione al corpus maturo della teoria dei kernel methods e delle Two-Sample Tests.

Questa connessione ha implicazioni per le garanzie statistiche: i risultati classici sul MMD (Gretton et al., 2012) garantiscono che il test ha potenza 1 asintoticamente contro qualsiasi alternativa non Gaussiana, e che il test statistico converge in distribuzione al suo valore atteso con velocità $O(1/\sqrt{N})$ per il Central Limit Theorem. La velocità di convergenza dell'approssimazione di SIGReg alla distribuzione target Gaussiana durante il training è quindi $O(1/\sqrt{N})$ nella dimensione del batch — più grande è il batch, più precisa è la stima di SIGReg come distanza dalla Gaussiana, e più efficace è la regolarizzazione.

17.4.1 Dimostrazione per stadi dell'identificabilità nel contesto JEPA: dal modello generativo ai fattori causali

La sezione 17.4 ha introdotto la connessione tra il teorema iVAE di Khemakhem et al. (2020) e le architetture JEPA, presentando la tesi principale — LeWorldModel come iVAE implicito — senza derivare la struttura della dimostrazione sottostante. Questa lacuna è significativa: la forza della connessione teorica dipende precisamente dall'identificazione di *quali condizioni del teorema iVAE* sono soddisfatte dall'architettura JEPA e in quali circostanze. La presente sottosezione sviluppa questa struttura dimostrativa in tre stadi, adattando la prova di Khemakhem et al. al contesto specifico dei sistemi JEPA con regolarizzazione SIGReg.

Il modello generativo JEPA come caso speciale dell'iVAE. Il punto di partenza è una riconcettualizzazione dell'architettura JEPA come modello generativo con variabile ausiliaria. Il modello iVAE canonico definisce una distribuzione congiunta:

$$p_{\{f, T, \lambda\}}(\mathbf{x}, \mathbf{z} \mid \mathbf{u}) = p_f(\mathbf{x} \mid \mathbf{z}) \cdot p_{\{T, \lambda\}}(\mathbf{z} \mid \mathbf{u}) / Z(\mathbf{u})$$

dove $f: \mathbb{Z} \rightarrow \mathcal{X}$ è la funzione di miscelazione (mixing function) invertibile, $p(\mathbf{x} \mid \mathbf{z}) = p_{\epsilon}(\mathbf{x} - f(\mathbf{z}))$ è la distribuzione del rumore di osservazione con $\mathbf{g}_{\epsilon}$ a supporto su $\mathcal{X}$, $p(\mathbf{z} \mid \mathbf{u})$ è la distribuzione dei fattori latenti condizionata all'ausiliario $\mathbf{u}$, e $Z(\mathbf{u}) = \int p_f(\mathbf{x} \mid \mathbf{z}) \cdot p_{\{T, \lambda\}}(\mathbf{z} \mid \mathbf{u}) d\mathbf{z}$ è la costante di normalizzazione. Nel contesto JEPA, la corrispondenza tra gli oggetti del modello iVAE e i componenti architetturali è:

La funzione di miscelazione f del modello iVAE corrisponde alla *funzione generativa* $g_{\theta} : \mathbb{Z} \rightarrow \mathcal{O}$ che è la pseudoinversa dell'encoder E_{θ} : $g_{\theta} = E_{\theta}^{\{-1\}}$ nel senso che $E_{\theta} \circ g_{\theta} \approx \text{Id}_{\mathbb{Z}}$. Nei sistemi JEPA non si addestra esplicitamente g_{θ} (non c'è ricostruzione), ma la sua esistenza è postulata dalla consistenza del modello generativo sottostante.

La distribuzione $p(z \mid u)$ della variabile latente condizionata all'ausiliario corrisponde alla distribuzione predittiva $p(z_{t+1} \mid a_t)$ — la distribuzione del successore latente condizionata all'azione. Quando il predittore pred_{φ} è affine in z_t con offset dipendente da a_t (la struttura AdaLN di LeWorldModel), questa distribuzione è Gaussiana con parametri dipendenti dall'azione:

$$p(z_{t+1} \mid z_t, a_t) = \mathcal{N}(f_{\varphi}(z_t, a_t), \sigma^2_{\varphi}(z_t, a_t) \cdot I_d)$$

Integrando su z_t rispetto a $p_{\text{SIGReg}}(z_t) \approx \mathcal{N}(0, I_d)$, si ottiene la distribuzione marginale $p(z_{t+1} \mid a_t)$ che è la distribuzione condizionale all'ausiliario $u = a_t$.

La variabile ausiliaria u del modello iVAE corrisponde all'azione a_t . L'azione è esogena rispetto alle variabili latenti (non è generata da un meccanismo che dipende dallo stato latente nel regime offline), e porta informazione discriminativa su quale transizione è avvenuta — esattamente il ruolo della variabile ausiliaria nel teorema iVAE.

Primo stadio: identificabilità della mixing function. Il primo stadio del teorema iVAE stabilisce che, se due modelli generativi ($g_{\theta}, p_{\{\mathbf{T}, \lambda\}}(z \mid a)$) e ($\tilde{g}_{\tilde{\theta}}, \tilde{p}_{\{\tilde{\mathbf{T}}, \tilde{\lambda}\}}(z \mid a)$) producono la stessa distribuzione osservazionale $p(o_{t+1} \mid a_t)$ per ogni valore dell'azione $a_t \in \mathcal{A}$, allora le due mixing function sono relate da una trasformazione: $g_{\theta} = \tilde{g}_{\tilde{\theta}} \circ h$ per qualche funzione invertibile $h : \mathbb{Z} \rightarrow \mathbb{Z}$.

La dimostrazione opera per contronominale. Supponiamo $p(o_{t+1} \mid a_t) = \tilde{p}(o_{t+1} \mid a_t)$ per tutto a_t . Poiché $p(o_{t+1} \mid a_t) = \int p(o_{t+1} \mid z') \cdot p(z' \mid a_t) dz'$ e $p(o_{t+1} \mid z') = p_{\varepsilon}(o_{t+1} - g_{\theta}(z'))$, la condizione di uguaglianza implica che le misure push-forward $(g_{\theta})^* p(z' \mid a_t)$ e $(\tilde{g}_{\tilde{\theta}})^* \tilde{p}(z' \mid a_t)$ coincidono su $\mathcal{O}$ per ogni valore di a_t . Poiché g_{θ} e $\tilde{g}_{\tilde{\theta}}$ sono diffeomorfismi invertibili (condizione necessaria per l'identificabilità), si può definire $h = g_{\theta}^{\{-1\}} \circ \tilde{g}_{\tilde{\theta}} : \mathbb{Z} \rightarrow \mathbb{Z}$, che è un diffeomorfismo. La tesi è $g_{\theta} = \tilde{g}_{\tilde{\theta}} \circ h$, ovvero $\tilde{g}_{\tilde{\theta}} = g_{\theta} \circ h^{\{-1\}}$, che segue direttamente dalla definizione di h .

Nel contesto JEPA, questo stadio garantisce che se l'encoder E_{θ} e un encoder alternativo $\tilde{E}_{\tilde{\theta}}$ producono la stessa distribuzione delle osservazioni future sotto le stesse azioni, allora i due spazi latenti sono connessi da un diffeomorfismo. Questo non è ancora identificabilità in senso forte — il diffeomorfismo h può essere arbitrariamente non lineare — ma stabilisce che la

struttura delle rappresentazioni è determinata dalla distribuzione osservazionale, non da scelte arbitrarie dell'architettura.

Secondo stadio: identificabilità component-wise sotto la condizione di sufficienza. Il secondo stadio impone la condizione strutturale sulla distribuzione $p(z \mid u)$ che restringe il diffeomorfismo h a essere *elementwise* — ossia una trasformazione che agisce separatamente su ciascuna coordinata dello spazio latente. Il teorema richiede che $p(z \mid u)$ appartenga alla famiglia esponenziale con statistiche sufficienti:

$$p(z \mid u) = Q_z(z) / Z(u) \cdot \exp(\sum_i \sum_j T_{ij}(z_i) \cdot \lambda_{ij}(u))$$

dove $T_{ij}(z_i)$ sono le statistiche sufficienti (funzioni della sola coordinata i -esima) e $\lambda_{ij}(u)$ sono i parametri naturali dipendenti dall'ausiliario. La condizione di sufficienza richiede che la matrice jacobiana $\Lambda(u) = [\lambda_{ij}(u)]_{i,j}$, valutata in $d + 1$ punti distinti $u_0, u_1, \dots, u_d$ nel supporto di $p(u)$, soddisfi:

$$\text{rango della matrice } [(\lambda_{ij}(u_1) - \lambda_{ij}(u_0)), \dots, (\lambda_{ij}(u_d) - \lambda_{ij}(u_0))]_{i,j} = d \cdot M_i$$

dove M_i è il numero di statistiche per la coordinata i -esima. Nel caso più semplice $M_i = 2$ (distribuzione Gaussiana con media e varianza parametrizzate dall'azione), la condizione richiede che la matrice $2d \times d$ delle differenze dei parametri abbia rango d .

Il nucleo della dimostrazione del secondo stadio usa il fatto che, se h non è *elementwise*, esiste almeno un indice j e una funzione $h_j : \mathbb{Z} \rightarrow \mathbb{R}$ tale che h_j dipende da almeno due coordinate z_i e z_k con $i \neq k$. In tal caso, la distribuzione $\tilde{p}(z' \mid u) = (h^{-1})^* p(z \mid u)$ non appartiene alla famiglia esponenziale con statistiche $T_{ij}(z_i)$ dipendenti da un'unica coordinata. Dato che sia p sia $\tilde{p}$ devono appartenere alla stessa famiglia esponenziale (per la simmetria dell'ipotesi), si ottiene una contraddizione con la condizione di sufficienza dei parametri — che richiede che i parametri naturali siano informativi *per ciascuna* coordinata separatamente.

Per il contesto JEPa-LeWorldModel, la condizione di sufficienza si traduce nella richiesta che l'AdaLN del predittore — che produce scale $\gamma_i(a)$ e shift $\delta_i(a)$ dipendenti dall'azione per ciascuna coordinata i dello spazio latente — sia sufficientemente informativo: la matrice $[(\gamma_i(a_k), \delta_i(a_k)) - (\gamma_i(a_0), \delta_i(a_0))]_{i,k}$ valutata in d punti distinti di $\mathcal{A}$ deve avere rango d . Questa condizione è genericamente soddisfatta quando la rete che mappa l'azione a ai parametri AdaLN è non-degenere e lo spazio delle azioni è almeno d -dimensionale — entrambe condizioni verificate nelle architetture di controllo considerate.

Terzo stadio: unicità fino a trasformazione elementwise monotona. Il terzo stadio conclude la prova stabilendo che il diffeomorfismo

component-wise $h(z) = (h_1(z_1), \dots, h_d(z_d))$ è necessariamente monotono in ciascuna componente — ossia ciascuna $h_i : \mathbb{R} \rightarrow \mathbb{R}$ è una funzione monotona crescente o decrescente.

La dimostrazione si fonda sulla conservazione della struttura d'ordine della distribuzione. Poiché $p(z_i | u)$ ha supporto convesso (la retta reale, nel caso Gaussiano) e la statistica $T_{\{i\}}(z_i) = z_i$ (primo momento) è monotona crescente in z_i , l'unica trasformazione h_i che preserva la struttura di famiglia esponenziale è una trasformazione monotona. Se h_i non fosse monotona, esisterebbe un intorno in cui $h_i' < 0$, che invertirebbe la relazione d'ordine tra z_i e $h_i(z_i)$ — il che sarebbe incompatibile con la conservazione delle statistiche sufficienti.

La conseguenza operativa per JEPA è la seguente: i fattori latenti $z^{\{i\}}$ dell'encoder E_θ sono identificati fino a trasformazione elementwise monotona dei fattori latenti veri $s^{\{i\}}$ del mondo. Questo significa che per ogni coordinata i dello spazio latente, esiste una funzione monotona $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$ tale che $z^{\{i\}} = \varphi_i(s^{\{i\}})$. Le funzioni φ_i sono invertibili, quindi i fattori veri si recuperano come $s^{\{i\}} = \varphi_i^{-1}(z^{\{i\}})$ — che è esattamente l'operazione che un probing classifier lineare su $z^{\{i\}}$ esegue, se il range di φ_i è sufficientemente uniforme nell'intervallo di interesse.

Quantificazione delle condizioni di applicabilità nel contesto fisico. La prova iVAE è valida sotto ipotesi ideali — mixing function esattamente invertibile, distribuzione esattamente della famiglia esponenziale, azioni esattamente esogene. Nel contesto di un sistema JEPA su dati fisici reali, ciascuna di queste ipotesi è soddisfatta soltanto approssimativamente, e la qualità dell'identificabilità degrada proporzionalmente alla violazione di ciascuna ipotesi.

La condizione di invertibilità della mixing function g_θ è la più restrittiva: g_θ deve essere iniettiva, il che richiede che l'encoder non collassi distribuzioni diverse sulla stessa rappresentazione. SIGReg contribuisce a questa condizione imponendo che la distribuzione degli embedding sia diffusa su tutto $\mathbb{R}^d$ (una distribuzione puntuale sarebbe collassata), ma non garantisce l'iniettività point-wise. L'iniettività effettiva dell'encoder è quindi una condizione necessaria addizionale che non segue dalla regolarizzazione ma deve essere valutata empiricamente attraverso i test di probing causale del Capitolo 12.

La condizione di esogeneità delle azioni è violata nei dataset raccolti con policy non uniformi, dove la distribuzione delle azioni dipende dallo stato osservato. In questi casi, il percorso backdoor $a_t \leftarrow \pi(a_t | o_t) \leftarrow o_t \rightarrow z_t$ introduce una dipendenza spuria tra azioni e rappresentazioni che non corrisponde a una relazione causale ma a un artefatto della policy di raccolta dati. La correzione di questo problema richiede tecniche di off-policy debiasing (importance weighting, counterfactual regression) che non sono attualmente integrate nelle architetture JEPA esistenti — una lacuna aperta che il sotto-programma I (CV-JEPA) del Capitolo 20 dovrà affrontare.

La condizione di famiglia esponenziale è la meno vincolante in pratica, poiché la famiglia Gaussiana con parametri dipendenti dall'azione è un'approssimazione ragionevole per le distribuzioni predittive di sistemi di controllo fisico continuo. Le deviazioni dalla Gaussianità (distribuzione multi-modale del successore in ambienti con contatti discontinui, come nel grasping) costituiscono le violazioni più significative — ed è precisamente in questi casi che l'identificabilità dei fattori causali degrada più marcatamente, come evidenziato dalla diminuzione del r^2 di probing nei task con dinamica discontinua rispetto ai task con dinamica continua.

17.4 La connessione con l'iVAE: identificabilità causale e regolarizzazione Gaussiana isotropa

Il risultato di identificabilità di Khemakhem, Kingma, Gower e Hyvärinen (2020) — noto come identifiable Variational Autoencoder (iVAE) — stabilisce condizioni sufficienti sotto cui i fattori latenti di un modello generativo sono identificabili, ossia recuperabili dalla distribuzione osservata a meno di permutazione elementwise e trasformazione monotona. Le condizioni del teorema principale di Khemakhem et al. sono tre: la distribuzione condizionale $p(x | z)$ è della forma esponenziale con statistiche sufficienti $s(x)$ e parametri $\lambda(z)$; la distribuzione $p(z | u) = p_u(z)$ è una distribuzione di famiglia esponenziale con parametri dipendenti da una variabile ausiliaria u (il segnale di supervisione debole); e il mapping dal parametro u al parametro della distribuzione di z è sufficientemente ricco (condizione di *sufficiency* dei parametri).

La connessione con il paradigma JEPa è sottile ma produttiva. In JEPa, la «variabile ausiliaria» u del teorema iVAE ha un candidato naturale: l'azione a_t . La distribuzione $p(z_{t+1} | a_t)$ — ottenuta integrando su z_t la distribuzione predittiva $\text{pred}_\varphi(z_t, a_t)$ — è una distribuzione condizionale all'azione che si sposta nello spazio latente in modo dipendente da a_t . Se questa distribuzione condizionale appartiene alla famiglia esponenziale con parametri $\eta(a_t)$ variabili con l'azione (una condizione soddisfatta, per esempio, quando pred_φ è affine in z_t con coefficienti dipendenti da a_t attraverso l'AdaLN di LeWorldModel), allora il teorema iVAE garantisce l'identificabilità dei fattori latenti.

Formalmente, sia $\eta(a_t) : \mathcal{A} \rightarrow \mathbb{R}^d$ il parametro della distribuzione $p(z_{t+1} | z_t, a_t)$ trattata come variabile ausiliaria. La condizione di sufficienza dei parametri del teorema iVAE richiede che la matrice jacobiana $J_{\eta} = \partial \eta / \partial a$ valutata in d punti distinti nel supporto di $p(a)$ abbia rango d — ossia che le azioni portino informazione indipendente su ciascuno dei d fattori latenti. Nei task di controllo continuo considerati in LeWorldModel (PushT, OGBench-Cube, TwoRoom, Reacher), lo spazio delle azioni è continuo e k -dimensionale con $k \geq 2$, e la condizione di rango è genericamente soddisfatta.

quando il numero di azioni è almeno d (il numero di fattori latenti), che è tipicamente verificato nelle traiettorie reali di controllo.

Quando la condizione iVAE è soddisfatta, il fattore latente $z^{\{i\}}$ apprende a rappresentare una quantità *identificabile*: una funzione invertibile di uno dei fattori latenti veri dello stato del mondo. Questo è esattamente il tipo di rappresentazione che, nella terminologia della presente monografia (Capitolo 1), si chiama *causalmente disentangled*: ciascuna coordinata dello spazio latente corrisponde a un fattore causale del mondo che ha un'identità indipendente dagli altri fattori.

La combinazione di SIGReg (che impone la distribuzione marginale Gaussiana isotropa) e del training objective JEPA condizionato sull'azione (che fornisce la variabile ausiliaria $u = a_t$) realizza quindi, almeno in linea di principio, le condizioni del teorema iVAE. La proposizione che emerge da questa analisi è la seguente: LeWorldModel è un iVAE implicito, nel senso che il suo obiettivo di training ottimizza implicitamente le condizioni di identificabilità causale, anche se nessuna delle due componenti (il predittore e il regolarizzatore) è stata progettata esplicitamente con questa finalità. Questa è la connessione teorica più profonda che emerge dalla sintesi del paradigma JEPA con il Causal Representation Learning: l'identificabilità causale non è una proprietà aggiuntiva da imporre, ma emerge naturalmente dalla combinazione di un predittore condizionato all'azione e di un regolarizzatore che impone una distribuzione a fattori indipendenti.

Un caveat importante va segnalato: il teorema iVAE garantisce l'identificabilità *fino a trasformazione elementwise monotona*, non l'identità perfetta con i fattori latenti veri. In pratica, questo significa che lo spazio latente di LeWorldModel non impara necessariamente la posizione assoluta del blocco in pixel, ma apprende una funzione monotona di tale posizione — una funzione che può essere invertita da un probing classifier lineare (come mostrato nella Tabella 1 del paper, dove LeWM raggiunge $r^2 = 0.986$ per la posizione del blocco nel task PushT). Questa corrispondenza tra la teoria dell'identificabilità e i risultati empirici di probing è una conferma qualitativa della connessione teorica derivata.

17.5 seq-JEPA e l'azione come variabile interventale: il caso dell'equivarianza sequenziale

L'architettura seq-JEPA di Ghaemi, Muller e Bakhtiari offre una prospettiva distinta ma complementare sulla formalizzazione dell'intervento nel paradigma JEPA. In seq-JEPA, l'azione non è un segnale di controllo nel senso del reinforcement learning, ma la *trasformazione relativa* tra due viste sequenziali di uno stesso oggetto o scena — per esempio la rotazione relativa tra due

rendering di un oggetto 3D in 3DIEBench, o la trasformazione di data augmentation relativa tra due crop di un'immagine.

La struttura formale dell'architettura seq-JEPA è la seguente. Siano $\{x_i\}_{i=1}^{M+1}$ una sequenza di viste generate dallo stesso campione x attraverso trasformazioni $\{t_i\}_{i=1}^{M+1}$, e siano $\{a_i\}_{i=1}^M$ le azioni (trasformazioni relative) che portano da x_i a x_{i+1} . L'encoder f_θ produce rappresentazioni $\{z_i\}_{i=1}^M$; l'aggregatore sequenziale g_φ (un transformer encoder con token [AGG]) produce la rappresentazione aggregata:

$$z_{\text{AGG}} = g_\varphi((z_1, a_1), (z_2, a_2), \dots, (z_{M-1}, a_{M-1}), z_M) \in \mathbb{R}^d$$

e il predittore MLP h_ψ stima la rappresentazione della vista $M+1$ come $\hat{z}_{M+1} = h_\psi(z_{\text{AGG}}, a_M)$. L'obiettivo di addestramento è la cosine similarity negativa:

$$\mathcal{L}_{\{\text{seq-JEPA}\}} = 1 - (\hat{z}_{M+1} / \|\hat{z}_{M+1}\|_2) \cdot (\text{sg}(z_{M+1}) / \|\text{sg}(z_{M+1})\|_2) \quad .$$

dove $\text{sg}(\cdot)$ è il stop-gradient che previene il collasso, con il peso del target encoder aggiornato tramite Exponential Moving Average (EMA) dei pesi di f_θ .

L'interpretazione causale di seq-JEPA è la seguente: l'azione a_i è una variabile interventale genuina nel senso di Pearl, non soltanto una variabile osservata. La trasformazione relativa che porta da x_i a x_{i+1} — sia essa una rotazione rigida in $\text{SO}(3)$, un crop, una trasformazione di colore, o uno spostamento saccadico — è prodotta dall'ambiente esterno (il processo di data augmentation, la cinematica del robot, il movimento oculare), non dall'osservazione passiva del sistema. In termini di DAG, ciò significa che la freccia causale va dall'ambiente all'azione (l'ambiente seleziona la trasformazione), non dall'osservazione all'azione — esattamente la struttura di un intervento $\text{do}(a_i = a)$ nel senso di Pearl.

Questo permette di interpretare l'obiettivo di addestramento di seq-JEPA come la minimizzazione della distanza tra la previsione della rappresentazione della vista $M+1$ sotto l'azione interventale a_M e la rappresentazione effettiva prodotta dall'encoder quando esposto alla vista $M+1$:

$$\mathcal{L}_{\{\text{seq-JEPA}\}} \approx d(\hat{z}_{M+1}, z_{M+1}) \text{ con } \hat{z}_{M+1} = h_\psi(z_{\text{AGG}}, \text{do}(a_M))$$

La proprietà emergente che distingue seq-JEPA da un semplice predittore condizionato è la *separazione architetturale* tra la rappresentazione equivariante $\{z_i\}$ e la rappresentazione invariante z_{AGG} . Gli esperimenti di ablation di Ghaemi et al. mostrano che, quando il conditioning sull'azione viene rimosso dall'aggregatore e dal predittore, le rappresentazioni $\{z_i\}$ perdono l'equivarianza (il R^2 per la previsione della rotazione relativa cade da 0.71 a 0.29)

mentre la classificazione invariante rimane competitiva (87.36% contro 86.05% di top-1 accuracy su 3DIEBench). Questo risultato empirico ha un'interpretazione causale diretta: l'action conditioning nel predittore e nell'aggregatore forza il sistema a separare la variazione causata dall'azione (che deve essere catturata nella rappresentazione equivariante z_i) dalla variazione invariante all'azione (che deve essere catturata in z_{AGG}).

Questa separazione è la realizzazione computazionale del principio di *autonomia dei meccanismi causali* applicato alla percezione visiva: il meccanismo che genera la componente equivariante della rappresentazione (dipendente dall'azione) e il meccanismo che genera la componente invariante (indipendente dall'azione) sono autonomi — appresi da componenti architettureali distinte (encoder vs. aggregatore) che non si informano a vicenda durante il forward pass.

La dimostrazione quantitativa di questo principio è fornita dall'esperimento di path integration di seq-JEPA. Dato un aggregatore z_{AGG} calcolato su una sequenza di M azioni $a_1, \dots, a_{\{M-1\}}$ e la rappresentazione della vista M , il modello può predire l'azione cumulativa $a_{\{1,M\}}$ (la trasformazione relativa che porta dalla vista 1 alla vista $M+1$) con $R^2 = 0.81$ per $M = 3$ rotazioni, usando soltanto z_{AGG} e a_M come input. Quando l'azione viene rimossa dall'input del predittore di path integration (ablation «No Action Conditioning»), l' R^2 cade a valori prossimi allo zero — il che dimostra che l'informazione sull'azione è conservata in z_{AGG} in forma composizionale, non come semplice media delle rappresentazioni. Questa composizionalità dell'informazione interventale nell'aggregatore è la proprietà che distingue seq-JEPA da un semplice modello sequenziale predittivo: è un modello di integrazione delle azioni nel senso di una World Model causale.

17.6 V-JEPA 2.1 e la densità spaziale come prerequisito della localizzazione interventale

V-JEPA 2.1 di Mur-Labadia, Muckley, Bar et al. introduce una modifica dell'obiettivo JEPA che ha implicazioni causali profonde, indipendentemente dall'intenzione esplicita degli autori. La modifica centrale — la Dense Predictive Loss — estende la supervisione dell'obiettivo JEPA da soltanto i token mascherati a *tutti* i token (sia mascherati sia di contesto):

$$\mathcal{L}_{\{dense\}} = \mathcal{L}_{\{predict\}} + \mathcal{L}_{\{ctx\}}$$

con

$$\mathcal{L}_{\{predict\}} = (1/|\mathcal{M}|) \sum_{\{i \in \mathcal{M}\}} \|P_{-\varphi}(E_{-\theta}(x), \Delta_{-y})_i - sg(E_{-\theta}(y))_i\|_1$$

$$\mathcal{L}_{\{ctx\}} = (1/|\mathcal{C}|) \sum_{\{i \in \mathcal{C}\}} \lambda_i \|P_{-\varphi}(E_{-\theta}(x), \Delta_{-y})_i - sg(E_{-\theta}(y))_i\|_1$$

dove $\mathcal{M}$ è l'insieme dei patch mascherati, $\mathcal{C}$ l'insieme dei patch di contesto, e $\lambda_i = \lambda / \sqrt{d_{\min}(i, \mathcal{M})}$ è il peso dipendente dalla distanza minima di ciascun patch di contesto al mascheramento più vicino.

L'interpretazione causale della Dense Predictive Loss emerge dalla distinzione tra due tipi di token nel DAG della V-JEPA: i token di contesto (visibili, che fungono da cause dell'inferenza) e i token mascherati (da prevedere, che fungono da effetti dell'inferenza). Nel V-JEPA originale, la supervisione sui token di contesto era assente, il che creava un incentivo perverso: i token di contesto potevano degradare in «registro globale» (nel senso di Darcet et al., 2023) — segnali che aggregano informazione dall'intera scena per facilitare la previsione dei token mascherati, senza mantenere informazione localmente strutturata sulla propria posizione spaziale.

Questo degrado è, in termini causali, una violazione del principio di *localizzazione causale*: ogni token dovrebbe rappresentare causalmente il proprio patch della scena, non una combinazione di informazione proveniente da patch distanti. La Dense Predictive Loss ripristina la localizzazione causale imponendo che ciascun token di contesto sia prevedibile anche dalla propria posizione — il che forza l'encoder a mantenere informazione localmente strutturata in ciascun token.

Formalmente, si consideri il DAG esteso per V-JEPA 2.1: siano s_i le variabili latenti dello stato del mondo alla posizione spaziale i (per esempio la normale di superficie, la profondità, la categoria semantica), e siano $z_i^{\{x\}}$ e $z_i^{\{y\}}$ i token latenti prodotti dall'encoder per le due viste x (mascherata) e y (non mascherata) della stessa scena. Il meccanismo causale che porta $s_i \rightarrow z_i^{\{y\}}$ è il meccanismo di codifica del patch i . La Dense Predictive Loss impone che $P_{\varphi}(E_{\theta}(x), \Delta_y)_i \approx z_i^{\{y\}}$ per ogni i , inclusi i token di contesto — il che, nell'interpretazione causale, forza che il predittore impari la funzione causale $s_i \rightarrow z_i^{\{y\}}$ per ogni patch i , non soltanto per i patch mascherati.

Il risultato empirico che conferma questa interpretazione è la qualità delle feature dense prodotte da V-JEPA 2.1. La PCA delle feature prodotte dall'encoder di V-JEPA 2 mostra feature maps rumorose e frammentate; quella di V-JEPA 2.1 mostra strutture spazialmente coerenti dove parti semanticamente simili (le ruote di automobili diverse, le teste di cani diversi) sono mappate agli stessi componenti PCA. Questa coerenza semantica è esattamente la proprietà attesa da un encoder che ha appreso la funzione causale $s_i \rightarrow z_i$ per ogni patch i in modo localmente consistente: patch con lo stesso tipo di contenuto fisico mappano allo stesso punto nello spazio latente perché il meccanismo causale che li genera è lo stesso (modulo trasformazioni geometriche).

La connessione con le architetture di robotica è diretta. V-JEPA 2.1 con V-JEPA-2 AC in task di grasping robotico su un braccio Franka Panda mostra un miglioramento del 20% nel tasso di successo del grasping rispetto a V-JEPA 2. Gli autori attribuiscono questo miglioramento alla migliore codifica delle

informazioni di profondità (depth), che è esattamente una proprietà del tipo di localizzazione causale descritta: la profondità di un oggetto è una variabile causale locale, non globale — dipende dalla struttura 3D della scena nella regione spaziale del patch, non dall'intera scena.

17.7 Il programma unificante: conseguenze architetture verificabili

La convergenza delle analisi delle tre architetture — LeWorldModel, seq-JEPA e V-JEPA 2.1 — attorno agli stessi principi causali permette di formulare il programma unificante in termini precisi. Il programma non è una mera ricognizione di analogie: è una serie di affermazioni con conseguenze architetture verificabili, ossia affermazioni che predicono differenze di comportamento tra architetture che soddisfano i principi causali e architetture che non li soddisfano.

La prima affermazione è che la *distribuzione marginale Gaussiana isotropa degli embedding* — realizzata da SIGReg in LeWorldModel — non è soltanto un meccanismo anti-collasso ma una condizione sufficiente per l'applicabilità della regola di aggiustamento backdoor in forma chiusa. La previsione verificabile è che sistemi con SIGReg generalizzino meglio a nuove azioni non viste nel training rispetto a sistemi con regolarizzatori non-Gaussiani, a parità di dataset. Questa previsione non è stata ancora testata sistematicamente nella letteratura esistente — è una lacuna aperta che il programma unificante identifica come priorità sperimentale.

La seconda affermazione è che la *separazione architetture tra componente equivariante e componente invariante* — realizzata attraverso l'aggregatore sequenziale di seq-JEPA — produce rappresentazioni più disentangled causalmente rispetto alle architetture a singolo encoder, nella misura in cui il disentanglement causale richiede la separazione tra ciò che varia con l'azione e ciò che è invariante all'azione. La previsione verificabile è che sistemi seq-JEPA, applicati al dominio del controllo robotico offline, apprendano rappresentazioni con probing accuracy più alta sui fattori causalmente separabili (posizione vs. orientamento vs. forma) rispetto a sistemi JEPA senza sequenza. Anche questa previsione non è stata testata nel paper originale, che si concentra sui benchmark di classificazione e equivarianza, non sui task di controllo.

La terza affermazione è che la *Dense Predictive Loss di V-JEPA 2.1* produce feature localmente causali che sono prerequisite per la pianificazione degli interventi in task spazialmente risolti (grasping, navigazione, segmentazione). La previsione verificabile è che sistemi con Dense Predictive Loss pianifichino in modo più accurato in task che richiedono localizzazione precisa nello spazio 3D rispetto a sistemi con obiettivo soltanto sui token mascherati. Questa previsione è parzialmente confermata dai risultati di V-JEPA 2.1 sul grasping (+20% su Franka Panda) e sulla navigazione (10× speedup su Tartan Drive con minore

ATE), anche se la separazione del contributo della Dense Predictive Loss dagli altri contributi (scaling, distillazione, multimodal tokenizer) non è controllata separatamente.

La convergenza di queste tre affermazioni costituisce il nucleo del programma unificante. Esso afferma che le tre proprietà — distribuzione marginale Gaussiana isotropa, separazione equivariante-invariante, densità spaziale della supervisione — non sono accidentalmente presenti nelle architetture più performanti, ma sono manifestazioni di una struttura causale sottostante: quella del DAG generativo esplicito che connette le variabili latenti del modello alle variabili latenti del mondo attraverso il do-calcolo di Pearl. Il programma unificante predice che sistemi che soddisfano tutte e tre le proprietà simultaneamente supereranno sistemi che ne soddisfano soltanto alcune su task che richiedono generalizzazione interventistica genuina — ossia task in cui la distribuzione delle azioni al test time differisce sistematicamente dalla distribuzione al training time, che è esattamente la struttura del problema di controllo in ambienti fisici aperti.

Capitolo 18 — L'intelligenza incarnata e verificabile come ideale regolativo

18.1 L'ideale regolativo e la sua funzione nel trittico: raccordo con le acquisizioni formali del Capitolo 17

Un ideale regolativo, nel senso kantiano del termine, non è un oggetto da raggiungere ma una direzione verso cui tendere: fornisce orientamento senza pretendere di essere pienamente realizzato nell'esperienza. La nozione di «intelligenza incarnata e verificabile» che guida la presente monografia — e che percorre, in forma latente, anche le due monografie precedenti del trittico — ha precisamente questa struttura: non descrive un sistema già esistente, ma articola le proprietà che un sistema agentivo dovrebbe soddisfare perché sia non soltanto capace ma *affidabile* nel senso pieno della parola, ossia verificabile dall'esterno nella qualità delle proprie rappresentazioni, controllabile nel proprio comportamento attraverso interventi esogeni, e radicato in un'ontologia del mondo fisico che non si riduce a pattern statistici.

L'ideale si compone di tre elementi distinti che si integrano senza collassare l'uno nell'altro. Il primo elemento è l'*incarnazione*: il sistema deve agire in un ambiente fisico continuo, ricevere osservazioni sensoriali non pre-processate e produrre azioni che hanno effetti fisici — non soltanto produrre previsioni simboliche su input pre-segmentati. Il secondo elemento è la *verificabilità*: le rappresentazioni interne del sistema devono essere sufficientemente strutturate

da permettere a un osservatore esterno di determinare cosa il sistema ha appreso, attraverso tecniche di probing, visualizzazione e test controllati — non soltanto valutare il comportamento esterno su un benchmark fisso. Il terzo elemento è la *coerenza causale*: il sistema deve modellare il mondo in termini di relazioni di dipendenza interventale, non soltanto correlazionali, in modo da generalizzare a situazioni nuove attraverso il ragionamento controfattuale invece della semplice interpolazione statistica.

Questi tre elementi sono presenti, in gradi diversi di compiutezza, nelle tre architetture analizzate nei capitoli precedenti e nel Capitolo 17: LeWorldModel realizza la coerenza causale attraverso SIGReg e la struttura predittiva interventale; seq-JEPA realizza la verificabilità attraverso la separazione architetturale tra componente equivariante e componente invariante; V-JEPA 2.1 realizza l'incarnazione attraverso le feature dense localizzate e la loro applicazione diretta in task di manipolazione e navigazione robotica.

Prima di esaminare in quale misura questi sistemi si avvicinino all'ideale, è però necessario rendere conto con rigore di quale fondamento formale il Capitolo 17 abbia effettivamente edificato per ciascuna delle tre componenti — e di quali rimangano al livello di programma. Il Capitolo 17 ha prodotto acquisizioni asimmetriche per i tre pilastri, e questa asimmetria struttura le valutazioni normative del Capitolo 18.

La *coerenza causale* è la componente più solidamente fondata. Le sezioni 17.1–17.2 dimostrano che, sotto la condizione di correttezza del DAG JEPA — ovvero quando enc_θ cattura le variabili rilevanti di S_t con fedeltà sufficiente da rendere $p(S_t \mid z_t)$ concentrata — la distribuzione interventale $p(z_{t+1} \mid z_t, \text{do}(a_t = a))$ coincide con la distribuzione condizionale osservazionale $p(z_{t+1} \mid z_t, a_t)$. Questo è un'equivalenza formale, non un'approssimazione, che segue dalla struttura del DAG JEPA con azioni esogene, e che rende $\text{pred}_\varphi(z_t, a_t)$ un estimatore diretto della distribuzione interventale senza richiedere il do-calcolo come operazione separata. La fondazione è però condizionale: vale per la classe $\mathcal{F}^*$ identificata dalla sezione 19.7 — SCM con transizioni Lipschitz-continue, rumore gaussiano, e azioni esogene. Al di fuori di $\mathcal{F}^*$ la fondazione del Capitolo 17 non si applica, e la coerenza causale rimane aspirazionale.

La *verificabilità* è fondata in modo più condizionale ancora. La sezione 17.4 dimostra che, sotto le condizioni di identificabilità dell'iVAE (Khemakhem et al.), l'ottimizzazione congiunta di $\mathcal{L}_{\text{JEPA}} + \lambda \cdot \text{Reg}_{\text{SIG}}$ converge a un encoder i cui fattori latenti z_t sono in biiezione con i veri fattori causali S_t , a meno di una trasformazione elementwise. Questa biiezione è la preconditione formale per la verificabilità: se z_t è in corrispondenza biunivoca con S_t , qualsiasi tecnica di probing sui fattori latenti produce risultati che rispecchiano la struttura causale reale. La fondazione della verificabilità è dunque ben definita nel caso iVAE — ma richiede stazionarietà a tratti del processo e l'esistenza di una variabile ausiliaria con variabilità contestuale strutturata. In ambienti

non-stazionari, o in assenza di questa struttura ausiliaria, la verifica teorica di 17.4 non è applicabile, e la verificabilità degrada alle proprietà empiricamente osservate — il VoE test di LeWorldModel, le visualizzazioni t-SNE di seq-JEPA — senza garanzia di identificabilità.

L'incarnazione, infine, è la componente per cui il Capitolo 17 ha fornito il fondamento più circoscritto. La sezione 17.5 mostra che seq-JEPA, attraverso la componente equivariante R_φ che trasforma la rappresentazione in risposta ad a_t , tratta esplicitamente l'azione come variabile interventale nel DAG latente: lo stato aggiornato $z'_t = R_\varphi(z_t, a_t)$ è la realizzazione computazionale dell'intervento $\text{do}(a_t = a)$ nel latente. La sezione 17.6 mostra che V-JEPA 2.1, attraverso le feature dense localizzate, fornisce l'aggancio sensoriale necessario per la manipolazione fisica. Tuttavia queste realizzazioni coprono soltanto l'incarnazione nel dominio visuo-motorio con osservazioni RGB: non affrontano la propriocezione continua, il controllo di forza e contatto, la gestione di dinamiche non-rigide, né il ciclo percettivo-motorio in tempo reale richiesto dall'incarnazione nel suo senso pieno. Il Capitolo 17 ha prodotto i mattoni iniziali dell'incarnazione — action conditioning equivariante, feature dense localizzate, planning nel latente — ma non la struttura completa del ciclo sensomotorio.

Il Capitolo 18 opera dunque a partire da questa fondazione asimmetrica: la coerenza causale è derivata per $\mathcal{F}^+$; la verificabilità è derivata sotto condizioni iVAE; l'incarnazione è parzialmente realizzata per il dominio visuo-motorio. L'ideale regolativo non è un miraggio, ma nemmeno un risultato già conseguito: è una direzione formalmente articolata il cui avanzamento è misurabile — estendendo $\mathcal{F}^+$, rilassando le condizioni iVAE, integrando sensori propriocettivi nel framework predittivo — e il Capitolo 18 esamina le tre architetture alla luce di questo avanzamento misurabile.

18.2 LeWorldModel come caso di studio: la verifica fisica come capacità emergente

La valutazione più sistematica della verificabilità di un sistema JEPA disponibile nella letteratura esaminata è quella condotta da Maes, Le Lidec, Scieur, LeCun e Balestriero su LeWorldModel nel contesto del Violation-of-Expectation (VoE) framework. Il VoE framework, mutuato dalla psicologia dello sviluppo — dove misura la comprensione della fisica intuitiva nei neonati attraverso il tempo di fissazione su eventi fisicamente possibili vs. impossibili — viene qui trasferito alla valutazione computazionale: si misurano la «sorpresa» del modello (l'errore di previsione latente) su traiettorie non perturbate, visivamente perturbate (cambio di colore di un oggetto) e fisicamente perturbate (teletrasporto di un oggetto in una posizione casuale).

I risultati del VoE test su LeWorldModel — riportati nella Figura 10 del paper e nei dettagli dell'Appendice F.3 — mostrano che il modello assegna

consistentemente una sorpresa maggiore alle perturbazioni fisiche (teletrasporto) rispetto alle perturbazioni visive (cambio di colore) su tutti e tre gli ambienti testati (TwoRoom, PushT, OGBench-Cube), con un test t appaiato che dà $p < 0.01$ per il teletrasporto in tutti gli ambienti. La sorpresa per il cambio di colore del cubo in OGBench-Cube non è statisticamente significativa — il modello non si «sorprende» abbastanza del cambio di colore, probabilmente perché lo spazio latente di LeWorldModel cattura proprietà geometrico-strutturali (posizione, forma) più che proprietà fotometriche (colore), una conseguenza attesa dalla natura del task di previsione del successore latente.

Questo risultato è rilevante per il programma unificante per diverse ragioni. In primo luogo, il VoE test è una forma di verifica *causale*: misura non soltanto la correttezza delle previsioni del modello su traiettorie osservabili, ma la sua sensibilità a violazioni della fisica causale — eventi che sono visivamente continui ma causalmente discontinui (il teletrasporto viola la continuità della traiettoria nello spazio fisico, non soltanto nel pixel space). Un modello che apprende soltanto correlazioni statistiche potrebbe non essere sensibile a questa distinzione, perché la continuità pixelwise e la continuità causale sono statisticamente correlate ma non equivalenti.

In secondo luogo, il pattern differenziale di sorpresa tra perturbazioni fisiche e perturbazioni visive è una firma qualitativa che distingue i modelli con struttura causale da quelli senza. Un modello che modella relazioni di dipendenza interventale dovrebbe essere più sensibile a violazioni delle invarianti fisiche (quantità di moto, continuità spaziale) che a variazioni delle proprietà fotometriche (colore), perché le prime sono vincolate dalla struttura del DAG causale (la posizione al tempo $t+1$ è causalmente dipendente dalla posizione al tempo t e dall'azione, non dal colore) mentre le seconde non lo sono. I risultati di LeWorldModel sono qualitativamente consistenti con questo pattern.

Terzo, e forse più significativo: questa capacità di discriminare la fisica causale emerge *senza supervisione esplicita sulla fisica*. LeWorldModel è addestrato soltanto sull'obiettivo predittivo latente e sulla regolarizzazione SIGReg — non su alcun segnale che espliciti cosa conta come «violazione fisica». Il fatto che la verifica causale emerga come proprietà del modello addestrato è un'istanza del fenomeno che il Capitolo 5 ha chiamato «comprensione emergente della fisica»: la struttura predittiva dell'obiettivo JEPA, combinata con la regolarizzazione appropriata, crea un incentivo implicito per l'encoder a separare le variabili fisicamente invarianti (posizione, velocità, forma) da quelle contingenti (colore, illuminazione), anche senza segnali di supervisione su questa distinzione.

Un aspetto dell'analisi di Maes et al. che merita un'attenzione particolare è la visualizzazione dello spazio latente attraverso il decoder. Il paper mostra (Figura 8) che, nel corso del training, lo spazio latente di LeWorldModel sviluppa progressivamente la capacità di supportare la decodifica di immagini visivamente plausibili: a 0 step di training, le immagini decodificate sono

uniformi; a 10k-40k step, emergono «slow features» (nel senso di Sobal et al., 2022) — strutture di bassa frequenza che corrispondono alle componenti più lentamente varianti nel tempo della scena; a 200k step, le immagini decodificate sono visivamente simili alle immagini reali. Questa traiettoria di apprendimento è consistente con la teoria dell'apprendimento delle slow features (Wiskott e Sejnowski, 2002): l'obiettivo predittivo latente favorisce l'apprendimento di rappresentazioni che cambiano lentamente nel tempo, le quali corrispondono alle variabili causalmente stabili (posizione, forma) piuttosto che a quelle instabili (intensità luminosa, artefatti di compressione).

La visualizzazione t-SNE dello spazio latente (Figura 9 del paper) per il task PushT mostra che la struttura dello spazio latente preserva le relazioni spaziali della griglia di stati: stati fisicamente vicini (stessa posizione dell'agente, orientamenti simili del blocco) sono proiettati in modo vicino nello spazio latente 2D, con una relazione quasi-continua che suggerisce che lo spazio latente apprende una varietà differenziabile che approssima lo spazio degli stati fisici. Questa proprietà è il prerequisito geometrico per la pianificazione nel latente: il Cross-Entropy Method può ottimizzare sequenze di azioni nello spazio latente perché la funzione di costo (distanza latente dal goal) è differenziabile e non ha barriere artificiali causate da discontinuità dello spazio di embedding.

La verificabilità di LeWorldModel è quindi una verifica *del tipo emergente*: non è imposta dall'architettura come vincolo rigido, ma emerge come conseguenza della struttura dell'obiettivo JEPA e della regolarizzazione SIGReg. Questo tipo di verificabilità emergente è distinto dalla verificabilità formale imposta da constraint come quelli di Causal VAE (Yang et al.) o iVAE (Khemakhem et al.), dove le proprietà causali sono garantite per costruzione. Le garanzie di verificabilità di LeWorldModel sono più deboli — empiriche piuttosto che dimostrative — ma anche più robuste rispetto a perturbazioni architetturali minori, perché non dipendono da assunzioni strutturali rigide sulla forma del DAG generativo.

18.3 seq-JEPA e la percezione sequenziale incarnata: dal saccadico alla navigazione

La dimensione dell'incarnazione che seq-JEPA introduce nel paradigma JEPA è di tipo diverso rispetto a quella realizzata da LeWorldModel e V-JEPA 2.1. LeWorldModel è incarnato nel senso del controllo motorico (agisce in ambienti fisici a bassa dimensionalità); V-JEPA 2.1 è incarnato nel senso della percezione densa (codifica la struttura spaziale tridimensionale della scena). seq-JEPA, invece, è incarnato nel senso della *percezione attiva*: il sistema integra una sequenza di osservazioni ottenute attraverso azioni di selezione del punto di fissazione (saccadi simulati), e apprende rappresentazioni che codificano la posizione relativa di ciascuna vista nel contesto della sequenza.

Il paradigma della percezione attiva — elaborato nella tradizione computazionale da Bajcsy (1988), Aloimonos et al. (1988) e Ballard (1991), e riconsiderato recentemente in chiave dell'*embodied cognition* — afferma che la percezione non è un processo passivo di ricezione di segnali, ma un ciclo attivo di selezione dell'informazione attraverso movimenti del corpo (oculari, cefalici, locomotori) guidati dagli obiettivi dell'agente. L'architettura seq-JEPA è, in questa tradizione, un tentativo di modellare computazionalmente il loop percezione-azione a livello della percezione elementare: il sistema non riceve l'intera immagine di una volta, ma integra progressivamente patch parziali nell'ordine determinato dalla sequenza di fissazioni.

La proposta «*predictive learning across saccades*» di Ghaemi et al. è implementata usando il dataset STL-10 con saliency maps generate dal modello DeepGaze IIE (Linardos et al., 2021), un modello pre-addestrato che predice la distribuzione delle fissazioni umane su immagini naturali. Ciascun campione di addestramento consiste in una sequenza di M patch 32×32 estratti dall'immagine 96×96 in posizioni selezionate dalla distribuzione di saliency (con Inhibition of Return per evitare la rivisitazione di aree già fissate), e seq-JEPA apprende a predire la rappresentazione della patch $M+1$ dall'aggregato z_AGG delle prime M patch condizionato sull'azione a_M (lo spostamento relativo tra il centro della patch M e il centro della patch $M+1$).

I risultati quantitativi sono notevoli nella misura in cui dimostrano che l'apprendimento attraverso saccadi simulate produce rappresentazioni di qualità paragonabile all'apprendimento con augmentazioni globali su immagini complete: la classificazione top-1 di seq-JEPA su STL-10 con $M_val = 4$ patch e saliency-driven sampling raggiunge 83.44%, comparata con 85.23% di SimCLR addestrato con augmentazioni globali sull'intera immagine. Considerando che seq-JEPA usa patch di risoluzione molto più bassa (32×32 contro 96×96) e non esegue augmentazioni globali, questo risultato indica che l'informazione estratta dalla sequenza di fissazioni è quasi equivalente all'informazione disponibile nell'immagine intera — un'istanza computazionale del fenomeno psicofisico noto come «*saccadic integration*»: la capacità del sistema visivo umano di costruire una rappresentazione coerente della scena da fissazioni sequenziali parziali.

La componente di path integration — la capacità di predire lo spostamento totale dalla vista 1 alla vista $M+1$ dato l'aggregatore z_AGG e l'ultima azione a_M — è l'aspetto che connette seq-JEPA più direttamente alla navigazione spaziale nel senso dell'*embodied cognition*. Gli esperimenti di Ghaemi et al. mostrano che il modello raggiunge $R^2 = 0.80$ per la previsione della posizione finale (coordinate pixel del centro dell'ultima patch) a partire dall'aggregatore con $M_val = 4$, e che questo R^2 si deteriora rapidamente quando le action embeddings vengono azzerate ($R^2 \approx 0.05$) — dimostrando che l'informazione di navigazione è codificata nelle action embeddings integrate dall'aggregatore, non soltanto nelle rappresentazioni visive.

Questo risultato ha un'interpretazione causale precisa: il sistema ha appreso a «integrare il cammino» nello spazio delle fissazioni, ossia a calcolare la posizione finale dall'accumulazione dei passi intermedi. Questa è la forma computazionale dell'integrazione del percorso (*path integration*) studiata nella navigazione animale — la capacità di mantenere una stima della posizione propria nell'ambiente aggiornando continuamente a partire dagli spostamenti relativi, senza riferimenti assoluti alla mappa globale. La sua realizzazione computazionale in seq-JEPA, attraverso il transformer aggregatore che accumula le action embeddings nella rappresentazione `z_AGG`, suggerisce che l'architettura seq-JEPA contiene in nucleo un modello di navigazione spaziale che potrebbe essere trasferito a task di navigazione robotica 3D con una modifica minima del training protocol.

Il confronto tra il tasso di classificazione con saliency-driven sampling (83.44%) e con random uniform sampling (79.85%) fornisce una misura quantitativa del vantaggio della percezione attiva orientata alla saliency: quasi 4 punti percentuali di classificazione top-1, attribuibili al fatto che le fissazioni guidate dalla saliency producono patch più informativi e meno ridondanti rispetto a fissazioni uniformemente casuali. Il meccanismo dell'*Inhibition of Return* contribuisce per ulteriori ~4 punti di differenza (79.85% senza IoR vs. 83.44% con IoR), confermando che la *diversità spaziale* delle fissazioni — ottenuta evitando la rivisitazione di aree già elaborate — è un componente critico della qualità dell'apprendimento percettivo.

18.4 V-JEPA 2.1 e la percezione densa come fondamento dell'interazione fisicamente radicata

V-JEPA 2.1 è, tra le tre architetture analizzate, quella che più direttamente realizza il terzo elemento dell'ideale regolativo — l'incarnazione nel senso del radicamento fisico. La Dense Predictive Loss produce feature maps spazialmente strutturate e semanticamente coerenti che sono prerequisite per i task di interazione fisica che richiedono comprensione della geometria 3D: stima della profondità, segmentazione semantica, inseguimento di oggetti in video, anticipazione delle interazioni.

I risultati sul task di stima della profondità (depth estimation) su NYUv2 — il più diretto proxy computazionale della comprensione della geometria 3D da immagini — illustrano il salto qualitativo tra V-JEPA 2 e V-JEPA 2.1: il RMSE cade da 0.642 a 0.307, una riduzione del 52% ottenuta con un probe lineare (un singolo strato lineare addestrato sulle feature frozen dell'encoder). Questo risultato supera persino DINOv3 con ViT a 7 miliardi di parametri (0.309 RMSE) — un modello che ha 3.5 volte più parametri — e AM-RADIO (0.340 RMSE) — il precedente stato dell'arte per modelli sotto i 2 miliardi di parametri. La superiorità di V-JEPA 2.1 su modelli di immagine puri per la

stima della profondità è particolarmente significativa perché la profondità è una proprietà intrinsecamente *dinamica*: cambia mentre l'agente o gli oggetti si muovono, e quindi beneficia dell'informazione temporale appresa attraverso l'obiettivo video che non è disponibile ai modelli addestrati soltanto su immagini statiche.

La visualizzazione qualitativa (Figura 10 del paper) mostra la differenza tra le depth maps prodotte da V-JEPA 2 e da V-JEPA 2.1: il primo produce mappe di profondità rumorose con struttura complessiva preservata ma bordi degli oggetti sfumati e inconsistenze locali; il secondo produce mappe nettamente definite con bordi precisi e coerenza locale elevata. Questa differenza è consistente con l'interpretazione causale del Capitolo 17: V-JEPA 2 apprende meccanismi di codifica globali (la patch dipende da molte altre patch attraverso i «token registro»), mentre V-JEPA 2.1 apprende meccanismi di codifica locali (la patch dipende causalmente dalla propria posizione nella scena) — e la profondità di un oggetto è una proprietà *locale*, determinata dalla struttura 3D nella regione del patch, non da quella dell'intera scena.

Il task di Video Object Segmentation (VOS) — dato il mask dell'oggetto nel primo frame, propagare il mask nei frame successivi usando soltanto la similarità delle feature — è un test particolarmente stringente della coerenza temporale delle rappresentazioni. V-JEPA 2.1 raggiunge 69.0 J&F su DAVIS-17 e 72.7 J&F su YouTube-VOS, secondi soltanto a DINOv3 (71.1 e 74.1 rispettivamente) e superiori a tutti i modelli con meno di 2 miliardi di parametri. La coerenza temporale delle feature di V-JEPA 2.1 è la conseguenza diretta dell'obiettivo video: il modello apprende che la stessa entità fisica (l'oggetto) deve essere rappresentata con feature simili attraverso frame consecutivi, anche quando cambia forma, posizione o illuminazione — una forma di *persistenza dell'oggetto* computazionale che è la controparte nella percezione artificiale della *permanenza dell'oggetto* studiata da Piaget nella cognizione infantile.

La task di Short-Term object-interaction Anticipation (STA) su Ego4D — dove il sistema deve predire dove, quando e con quale azione un agente umano interagirà con un oggetto nell'imminente futuro — porta V-JEPA 2.1 ad un mAP All di 7.71, un miglioramento relativo del 35% rispetto al precedente stato dell'arte (STAformer, 5.67 mAP) che incorporava componenti task-specific progettate ad hoc. Questo risultato è particolarmente significativo perché il task STA combina le tre proprietà dell'ideale regolativo: richiede incarnazione (il sistema deve comprendere le affordances fisiche degli oggetti e i vincoli del corpo umano), verificabilità (i risultati sono valutati con metriche precise di localizzazione spaziale, categorizzazione semantica e stima temporale) e coerenza causale (la previsione dell'interazione richiede di inferire l'*intenzione* dell'agente come variabile causale latente, non soltanto il pattern statistico delle interazioni passate).

18.5 La distribuzione dei risultati: cosa è già raggiungibile

L'esame sistematico dei risultati empirici di LeWorldModel, seq-JEPA e V-JEPA 2.1 permette di tracciare con precisione la mappa di ciò che è *già raggiungibile* nell'ambito dell'ideale regolativo dell'intelligenza incarnata e verificabile, distinguendolo da ciò che rimane speculativo o richiede sviluppi teorici e architetture non ancora disponibili.

Sul piano della *verificabilità emergente*, i sistemi JEPA attuali dimostrano tre proprietà già raggiungibili. La prima è la verificabilità attraverso probing: dati gli embedding di un sistema addestrato (LeWorldModel su PushT), un probing classifier lineare recupera la posizione dell'agente con $r^2 = 0.974$, la posizione del blocco con $r^2 = 0.986$ e l'orientamento del blocco con $r^2 = 0.902$ — valori che dimostrano che queste quantità fisiche sono linearmente codificate nello spazio latente e quindi verificabili da un osservatore esterno con un effort computazionale minimo. La seconda è la verificabilità attraverso decodifica: il decoder addestrato a posteriori sugli embedding di LeWorldModel produce immagini visivamente plausibili della scena anche a 200k step di training, confermando che lo spazio latente mantiene informazione sufficiente per la verifica visuale. La terza è la verificabilità attraverso test counterfattuale: il VoE test dimostra che il sistema distingue eventi fisicamente possibili da eventi fisicamente impossibili con specificità statistica ($p < 0.01$), il che equivale a una forma di verifica causale automatizzata.

Sul piano dell'*incarnazione effettiva*, i risultati dimostrano che i sistemi JEPA attuali sono già capaci di pianificazione efficace in ambienti fisici di media complessità. LeWorldModel pianifica task di manipolazione 2D (PushT, Reacher) e 3D (OGBench-Cube) con success rate competitivi rispetto ai baseline state-of-the-art, con un'efficienza computazionale di $48\times$ superiore a DINO-WM e con modelli da 15 milioni di parametri addestrabili su una singola GPU in poche ore. V-JEPA 2.1 guida un braccio robotico Franka Panda reale in task di grasping zero-shot con 80% di success rate usando MPC latente. Questi non sono risultati di laboratorio astratti: sono sistemi che operano su hardware fisico con osservazioni visive reali.

Sul piano della *coerenza causale*, i risultati di path integration di seq-JEPA ($R^2 = 0.80$ per la posizione finale a partire dall'aggregatore sequenziale) e i risultati di generalizzazione di LeWorldModel a nuove configurazioni iniziali (il success rate è calcolato su goal states campionati dalla distribuzione del dataset, non su configurazioni viste durante il training) suggeriscono che questi sistemi non memorizzano le traiettorie di training ma apprendono strutture generative che si generalizzano a situazioni nuove. Questo tipo di generalizzazione — che va oltre l'interpolazione statistica e si avvicina all'inferenza causale — è una delle proprietà più attese e più difficili da dimostrare nei sistemi di apprendimento.

18.6 Ciò che rimane speculativo: i confini attuali del paradigma

Accanto a ciò che è già raggiungibile, l'ideale regolativo dell'intelligenza incarnata e verificabile include proprietà che i sistemi attuali non realizzano ancora in modo robusto e che la letteratura disponibile non supporta con evidenze dimostrative.

La prima proprietà ancora speculativa è la *generalizzazione interventale a distribuzione shift sistematico*. I sistemi JEPA attuali dimostrano buona generalizzazione entro la distribuzione del dataset — goal states non visti ma campionati dalla stessa distribuzione delle traiettorie di training. Ma la generalizzazione a *shift sistematici* della distribuzione delle azioni — per esempio un sistema addestrato su traiettorie di esplorazione casuale che deve pianificare traiettorie ottimizzate, o un sistema addestrato in simulazione che deve operare nel mondo fisico — rimane una sfida aperta. Il sim-to-real gap, discusso nel Capitolo 8, è la manifestazione più concreta di questo problema: la distribuzione degli embedding di V-JEPA 2.1 su scene di simulazione differisce sistematicamente dalla distribuzione su scene reali a causa delle differenze nel rendering, nell'illuminazione e nella fisica, e questa differenza non è compensata dall'obiettivo di addestramento senza uno shift esplicito del dominio.

La seconda proprietà è la *verifica formale delle proprietà causali*. I test di verifica descritti — probing, VoE, visualizzazione t-SNE — sono test empirici che dimostrano la presenza di struttura causale nello spazio latente ma non la garantiscono formalmente. Un sistema che supera questi test potrebbe comunque «imbrogliare» (nel senso di appoggiarsi a correlazioni spurie che mimano la struttura causale nel regime di test ma non generalizzano a perturbazioni fuori distribuzione). La verifica formale richiederebbe la specificazione esplicita del DAG causale e la dimostrazione che il modello lo realizza — un programma che il Causal Representation Learning sta sviluppando (come discusso nel Capitolo 2 con riferimento a Causal VAE e iVAE) ma che non è ancora applicato a sistemi JEPA su larga scala in ambienti fisici reali.

La terza proprietà è la *pianificazione a lungo orizzonte in spazi fisici complessi*. LeWorldModel pianifica con orizzonte $H = 5$ step (corrispondenti a 25 timestep dell'ambiente per via del frame skip di 5), che è sufficiente per i task di manipolazione 2D e 3D considerati nel paper ma non per task che richiedono pianificazione su scale temporali più lunghe — task di navigazione in ambienti larghi, manipolazione di sequenze multi-step complesse, interazione con più oggetti in sequenza. L'accumulo di errori di previsione nell'auto-regressione latente limita l'efficacia della pianificazione a orizzonti più lunghi: l'errore di previsione al passo $t+k$ cresce tipicamente come $O(k^\alpha)$ con $\alpha > 0$, e per k sufficientemente grande, il segnale di previsione diventa troppo rumoroso per guidare l'ottimizzazione dell'azione. Le architetture gerarchiche per la

pianificazione a lungo orizzonte (World Models gerarchici, come discusso nel Capitolo 9) sono il candidato teorico più promettente per affrontare questa limitazione, ma la loro integrazione con il paradigma JEPA è un programma di ricerca ancora in corso di sviluppo, non una tecnologia consolidata.

La quarta proprietà sempre speculativa come le altre tre è la *composizionalità strutturale delle rappresentazioni*. I sistemi JEPA attuali apprendono rappresentazioni che sono funzioni continue e in genere non-lineari dei fattori fisici dello stato del mondo — la posizione dell'agente, l'orientamento del blocco, la profondità di una superficie. Ma non apprendono strutture composizionali esplicite nel senso della composizionalità sistematica richiesta dal ragionamento simbolico: la capacità di rappresentare «il blocco rosso sopra il cubo blu» come composizione di rappresentazioni del blocco rosso, del cubo blu e della relazione «sopra» — in modo tale che la stessa rappresentazione di «sopra» si combini produttivamente con qualsiasi coppia di oggetti. Questa proprietà è il principale punto di contatto tra il paradigma JEPA e i framework neuro-simbolici (NS-CL, discusso nel Capitolo 4), ed è il territorio in cui la distanza tra l'ideale regolativo e i sistemi attuali è più pronunciata.

Il riconoscimento esplicito di queste quattro limitazioni non è una concessione al pessimismo: è la preconditione per un programma di ricerca serio. Un ideale regolativo che si confonde con le capacità attuali cessa di svolgere la propria funzione normativa. La distinzione tra ciò che è raggiungibile e ciò che rimane speculativo permette di orientare il lavoro futuro verso le lacune reali invece che verso la conferma di proprietà già dimostrate — ed è in questa funzione orientativa che l'ideale regolativo dell'intelligenza incarnata e verificabile guadagna il proprio posto come concetto fondamentale del trittico JEPA.

Capitolo 19 — Inventario dei limiti strutturali residui: impossibilità, confini epistemici e lacune non contingenti

19.1 La funzione epistemica dei limiti in un programma di ricerca maturo

Ogni programma di ricerca scientifico serio si misura non soltanto con i propri successi dimostrativi ma con la chiarezza con cui traccia i confini di ciò che non può raggiungere. In fisica teorica questo compito è svolto dai teoremi di impossibilità: il teorema di Bell sul realismo locale, il teorema di Gödel sull'incompletezza, il teorema di Arrow sull'impossibilità della funzione di benessere sociale. Nell'apprendimento automatico, la tradizione dei teoremi di No Free Lunch e dei risultati di non-identificabilità svolge una funzione analoga — non per scoraggiare la ricerca ma per disciplinarla, impedendo che le

aspettative vengano sistematicamente ancorate a proprietà che per ragioni strutturali non possono essere garantite.

La presente monografia — e il trittico di cui fa parte — ha costruito un edificio teorico che connette il paradigma JEPA all'inferenza causale, all'incarnazione robotica e all'allineamento normativo. Il Capitolo 18 ha già distinto tra ciò che è già raggiungibile e ciò che rimane speculativo nel contesto delle tre architetture analizzate. Il presente capitolo approfondisce questa distinzione fino al livello dei limiti *strutturali* — quelli che non dipendono dall'insufficienza attuale di dati, di potenza computazionale o di ingegnosit  architetture, ma dalla geometria stessa del problema che si vuole risolvere. Questa distinzione non   un esercizio di pessimismo:   la base su cui un programma di ricerca onesto pu  costruire le proprie priorit  senza sprecare risorse nell'inseguimento di risultati che un'analisi attenta avrebbe mostrato irraggiungibili.

I limiti che si documentano in questo capitolo si collocano su tre piani distinti. Il primo   il piano dell'identificabilit  matematica: ci sono propriet  degli spazi latenti che non possono essere recuperate dalla distribuzione osservativa senza assunzioni strutturali forti, e queste assunzioni non sono sempre verificabili dai dati. Il secondo   il piano della supervisione epistemica: ci sono regimi di complessit  in cui nessun segnale di supervisione umana — n  ordinario n  complementare n  ricorsivo —   sufficiente a garantire l'allineamento del sistema, e questo confine non   un limite contingente della tecnologia attuale ma un confine strutturale dell'epistemologia della preferenza. Il terzo   il piano della verifica formale: i metodi di theorem proving per sistemi di controllo, sviluppati nella tradizione della verifica deuttiva con strumenti come Why3, offrono garanzie formali che i sistemi JEPA non possono ancora produrre per la propria struttura interna, e la distanza tra le garanzie producibili con metodi formali e le garanzie producibili con metodi empirici   una lacuna strutturale con conseguenze dirette per la sicurezza dei sistemi deployati.

19.2 Il problema dell'identit  numerica dei concetti attraverso le modalit  percettive

Il Capitolo 4 ha introdotto il problema dell'identit  trans-modale come questione aperta: due rappresentazioni — una visiva e una linguistica — codificano *lo stesso* concetto se e soltanto se rispondono allo stesso modo agli stessi interventi. Questa definizione operativa era presentata come un approccio parziale al problema, non come la sua risoluzione. Il Capitolo 19 approfondisce perch  questa parzialit  non sia contingente.

Il problema dell'identit  numerica trans-modale ha la seguente struttura formale. Siano $\mathcal{M}_{\text{vis}}$ un modello visivo con spazio latente $\mathbb{Z}_{\text{vis}} \subseteq \mathbb{R}^d$ e $\mathcal{M}_{\text{ling}}$ un modello linguistico con spazio latente $\mathbb{Z}_{\text{ling}} \subseteq \mathbb{R}^{d'}$. Un concetto C   rappresentato da una regione $\mathcal{R}_{\text{vis}}(C) \subseteq \mathbb{Z}_{\text{vis}}$ nel modello visivo e da una regione $\mathcal{R}_{\text{ling}}(C) \subseteq \mathbb{Z}_{\text{ling}}$ nel modello linguistico. La domanda  : esiste una

corrispondenza biunivoca $f_C : \mathcal{R}_{\text{vis}}(C) \rightarrow \mathcal{R}_{\text{ling}}(C)$ tale che per ogni coppia $(z_{\text{vis}}, z_{\text{ling}})$ con $z_{\text{ling}} = f_C(z_{\text{vis}})$ valgano le stesse relazioni di dipendenza interventale?

L'approccio causale stabilisce la condizione necessaria per l'identità: la distribuzione interventale $P(z_{\text{vis}} \mid \text{do}(C = v))$ nel modello visivo deve corrispondere — attraverso f_C — alla distribuzione interventale $P(z_{\text{ling}} \mid \text{do}(C = v))$ nel modello linguistico, per ogni valore v del concetto. Ma questa condizione non è sufficiente a garantire l'identità numerica, per una ragione che Quine ha chiamato *inscrutabilità del riferimento*: due sistemi possono rispondere allo stesso modo agli stessi interventi su concetti etichettati con lo stesso nome pur rappresentando *entità ontologicamente distinte*. Un sistema visivo che chiama «rosso» una certa regione dello spazio latente e un sistema linguistico che chiama «rosso» un'altra regione possono produrre distribuzioni interventali identiche sotto perturbazioni di stimoli standard, e tuttavia le due regioni potrebbero corrispondere a proprietà fisicamente diverse — la riflettanza spettrale nel caso visivo, la membership in una categoria semantica nel caso linguistico.

Il problema dell'inscrutabilità non è risolvibile attraverso il solo allineamento delle distribuzioni interventali perché queste distribuzioni sono misurate su insiemi di stimoli finiti, mentre le categorie concettuali sono definite su insiemi potenzialmente infiniti. Due sistemi con rappresentazioni diverse possono essere indistinguibili su qualsiasi insieme finito di stimoli pur differendo su stimoli fuori da quell'insieme. Formalmente: per qualsiasi insieme finito di test $\Omega = \{(x_i, \text{do}(C_j = v_k))\}_{i,j,k}$, esistono infiniti piani di mapping f_C che producono le stesse distribuzioni interventali su Ω ma divergono su stimoli non in Ω . Questo è un corollario diretto del teorema di non-identificabilità di Hyvarinen per i modelli di mistura non lineare applicato al caso del doppio modello — e il teorema vale indipendentemente dalla complessità dei modelli, non soltanto per i sistemi JEPA.

La conseguenza pratica per il programma di ricerca sull'allineamento multi-modale è che l'identità numerica dei concetti attraverso le modalità non può essere verificata empiricamente in modo esaustivo. Si può dimostrare la *compatibilità funzionale* — due sistemi producono le stesse uscite per gli stessi input su un insieme di test rappresentativo — ma non l'identità. Questa distinzione è cruciale per il scalable oversight: un sistema di supervisione che verifica la compatibilità funzionale tra le rappresentazioni visive del sistema AMI e le rappresentazioni linguistiche del supervisore umano non garantisce che il sistema abbia internamente lo stesso concetto del supervisore, ma soltanto che si comporta come se ce l'avesse nell'insieme dei test considerati.

19.3 La non-identificabilità strutturale degli SCM latenti in ambienti fisici

Il Capitolo 2 ha derivato le condizioni di identificabilità dell'iVAE di Khemakhem et al. e del CausalVAE di Yang et al., mostrando che entrambi richiedono assunzioni strutturali forti per recuperare i fattori latenti causalmente interpretabili. Il Capitolo 17 ha mostrato che LeWorldModel realizza implicitamente le condizioni dell'iVAE attraverso la combinazione di SIGReg (distribuzione marginale Gaussiana isotropa) e action conditioning (la variabile ausiliaria $u = a_t$). Ma il teorema di identificabilità dell'iVAE garantisce il recupero dei fattori latenti a meno di permutazione e trasformazione elementwise monotona — non il recupero assoluto. In ambienti fisici reali, questa garanzia residua è insufficiente per diversi motivi strutturali che si documentano qui.

Il primo motivo è che il teorema iVAE assume che la variabile ausiliaria u (l'azione a_t nel caso di LeWorldModel) sia *univocamente associata* a ciascun item del dataset — ogni traiettoria ha le proprie azioni specifiche, e la variabilità delle azioni fornisce la diversità necessaria per la condizione di sufficienza del teorema. In ambienti fisici reali, le azioni sono spesso ridondanti: due sequenze di azioni diverse possono portare allo stesso stato, il che riduce la diversità informazionale delle azioni come variabili ausiliarie. Se la matrice jacobiana $\partial\eta/\partial a$ ha rango inferiore a d in regioni dello spazio degli stati dove le azioni sono ridondanti (per esempio in prossimità dei goal states in task di manipolazione, dove molte sequenze di azioni producono lo stesso grasping finale), la condizione di sufficienza del teorema non è soddisfatta in quelle regioni, e i fattori latenti corrispondenti non sono identificabili.

Il secondo motivo è la presenza di *confondenti non osservati* negli ambienti fisici reali — variabili che influenzano sia le osservazioni sia le azioni ma non sono misurate dal sistema. Le proprietà fisiche degli oggetti — attrito, elasticità, massa — variano tra oggetti visivamente simili e influenzano le dinamiche degli oggetti stessi, ma non sono direttamente misurate dagli encoder visivi. Un sistema JEPA addestrato su dati con variabilità nelle proprietà fisiche degli oggetti senza accesso a misurazioni dirette di queste proprietà produrrà fattori latenti che confondono le proprietà fisiche non osservate con le proprietà geometriche osservate — una forma di mistura che l'iVAE non può separare senza variabili ausiliarie specifiche per le proprietà fisiche non osservate.

Il terzo motivo — il più fondamentale — è che l'identificabilità causale garantita dall'iVAE è *locale*, nel senso che dipende dall'invertibilità della matrice jacobiana $\partial\eta/\partial a$ in ogni punto del supporto della distribuzione. In spazi degli stati ad alta dimensionalità — come quelli degli ambienti fisici complessi — questa matrice può avere rango deficitario in sottospazi del supporto che corrispondono esattamente alle configurazioni più interessanti (stati vicini al goal, configurazioni di contatto, stati di biforcazione dinamica). Il teorema di identificabilità non garantisce nulla su questi sottospazi di degenerazione, e sono esattamente quelli in cui la qualità delle rappresentazioni causali è più critica per il controllo.

Formalmente, sia $J(a, z) = \partial \eta(a) / \partial a \in \mathbb{R}^{d \times k}$ la jacobiana delle statistiche sufficienti rispetto all'azione. La condizione di sufficienza del teorema iVAE richiede che esistano $d+1$ punti $\{a^{\{0\}}, a^{\{1\}}, \dots, a^{\{d\}}\}$ nel supporto di $P(a)$ tali che la matrice:

$$\Gamma = (\eta(a^{\{1\}}) - \eta(a^{\{0\}}), \eta(a^{\{2\}}) - \eta(a^{\{0\}}), \dots, \eta(a^{\{d\}}) - \eta(a^{\{0\}})) \in \mathbb{R}^{d \times d}$$

sia invertibile. Questa condizione è verificabile in linea di principio ma non in pratica: richiederebbe di conoscere $\eta(a)$ analiticamente, mentre in un sistema JEPA $\eta(a)$ è approssimato implicitamente dalla funzione $\text{pred}_\varphi(z, a)$, che è una rete neurale senza forma analitica chiusa. Il fallimento di questa condizione non è rivelato dall'obiettivo di addestramento — un sistema che minimizza la perdita JEPA non riceve nessun segnale che indichi la non-identificabilità dei fattori latenti nelle regioni dove $J(a, z)$ è degenere.

19.4 Il limite della recursive critique: l'asimmetria strong-to-weak come confine strutturale

Il framework del scalable oversight attraverso recursive self-critiquing — sviluppato da Wen, Lou, Lu, Yang et al. (2026) — ha dimostrato che la critica di ordine superiore è sistematicamente più facilmente valutabile della critica di ordine inferiore, e che questa gerarchia si mantiene ricorsivamente attraverso gli stadi C^1 , C^2 , C^3 . Questo risultato è epistemicamente promettente ma contiene un limite strutturale che il paper stesso documenta con precisione: nel regime strong-to-weak, in cui il modello critico è *più* capace del modello supervisionato, la recursive self-critique si rivela controproducente.

I risultati delle AI-AI experiments di Wen et al. mostrano che, quando i modelli supervisionati sono *più grandi* del modello critico da 7B (supervisionato da 14B a 72B), l'ordine superiore di critica migliora sistematicamente le prestazioni: il Performance Recovered (PR) cresce da C^1 a C^3 per modelli supervisionati di taglia 14B, 32B e 72B. Ma quando i modelli supervisionati sono *più piccoli* del modello critico (0.5B, 1.5B, 3B), la critique diretta C^1 produce risultati superiori a C^2 e C^3 — la self-critique del modello più debole non è utile e può persino fuorviare il modello critico più forte. Questa asimmetria non è un artefatto sperimentale: ha una spiegazione strutturale precisa.

Quando un modello più debole produce auto-critiche di ordine superiore — C^2 e C^3 — queste critiche non migliorano la qualità della supervisione rispetto alla critica diretta perché il modello non ha la capacità di generare insight genuinamente nuovi attraverso la meta-riflessione. La critica di ordine superiore del modello debole è una riorganizzazione verbale della critica di primo ordine, non una prospettiva qualitativamente più astratta. Il modello critico forte che legge questa auto-critica superiore del modello debole riceve un segnale più

confuso, non più chiaro — perché la critica del modello debole ha già introdotto distorsioni nella propria critica di primo ordine che si amplificano nell'auto-critica di secondo ordine.

Formalmente, sia B la capacità del modello base (supervisionato) e C la capacità del modello critico. Il Performance Recovered $PR(B, C, n)$ al livello n -esimo di critica è una funzione di tutte e tre le variabili. Wen et al. dimostrano empiricamente che:

$\partial PR / \partial n > 0$ quando $B > C$ (weak-to-strong: il supervisore è più debole del sistema)

$\partial PR / \partial n < 0$ quando $B < C$ (strong-to-weak: il supervisore è più forte del sistema)

Il confine $B = C$ è il punto critico dove la recursive critique è neutrale rispetto all'ordine. Questo confine ha un'interpretazione causale nella terminologia del Capitolo 1: la self-critique di ordine superiore produce un segnale di supervisione informativo soltanto quando il modello supervisionato ha già una struttura interna sufficientemente ricca da generare istanze genuine di meta-riflessione. Al di sotto di questa soglia, la self-critique è rumore epistemico mascherato da struttura sintattica.

La conseguenza per il programma di allineamento dei sistemi AMI è diretta e scomoda: il regime di maggiore interesse per il scalable oversight — quello in cui il sistema AMI supera le capacità del supervisore umano — è precisamente il regime $B > C$, che corrisponde al caso weak-to-strong. Ma il regime weak-to-strong è anche quello in cui i supervisori umani hanno meno capacità di verificare autonomamente la qualità della self-critique prodotta dal sistema — non possono distinguere se la critica di secondo ordine del sistema è un genuino approfondimento o una sofisticata razionalizzazione di un errore di primo ordine. Il framework della recursive critique è quindi più efficace esattamente dove è meno verificabile — un dilemma strutturale che non ha soluzione all'interno del paradigma della sola supervisione umana.

Gli esperimenti di Wen et al. sulle AI-AI tasks mostrano che anche nel regime weak-to-strong il meccanismo ha limiti precisi: i modelli della serie Qwen2.5-Instruct migliorano sistematicamente con la recursive critique su task matematici (MATH dataset), ma i guadagni sono concentrati negli intervalli di accuratezza moderata (0.3–0.7) e quasi assenti agli estremi — sia per task molto facili (accuratezza vicina a 1) che per task molto difficili (accuratezza vicina a 0). La spiegazione è analoga a quella dell'asimmetria strong-to-weak: la recursive critique è informativa soltanto quando le risposte iniziali contengono già sufficiente struttura da permettere una meta-riflessione produttiva. Per i task estremi, il sistema produce o risposte quasi sempre corrette (la critica non aggiunge nulla) o risposte troppo sistematicamente errate da permettere un miglioramento ricorsivo.

19.5 Il teorema di impossibilità dell'allineamento senza supervisione esaustiva

L'impossibilità di garantire l'allineamento senza supervisione esaustiva è il risultato negativo più fondamentale dell'intera letteratura del scalable oversight, e merita una trattazione formale che non si trova — nella sua completezza — in nessuno dei paper esaminati, ma che emerge dalla congiunzione di diversi risultati parziali.

Sia π il sistema AMI da allineare, r^* la funzione di ricompensa ideale che cattura le preferenze dell'utente su tutti gli input possibili, e r_ψ il modello di ricompensa addestrato su un dataset finito $D = \{(x_i, y_w^i, y_l^i)\}_{i=1}^N$ di confronti umani. Il problema dell'allineamento è trovare la policy π^* che massimizza $E\{x \sim p(x) [r^*(x, \pi^*(x))]\}$. Il problema del reward hacking è che r_ψ e r^* concordano su D ma non necessariamente su tutto il dominio $\mathcal{X} \times \mathcal{Y}$.

La discrepanza tra r_ψ e r^* è misurabile attraverso la distanza:

$$\Delta(r_\psi, r^*) = \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |r_\psi(x, y) - r^*(x, y)|$$

Il teorema seguente — derivabile dai principi generali dell'apprendimento statistico — stabilisce il confine strutturale.

Proposizione (impossibilità dell'allineamento senza supervisione esaustiva): Per qualsiasi policy π che ottimizza r_ψ al di sopra di una soglia $\theta > 0$, la degradazione dell'utilità rispetto a r^* è limitata inferiormente da:

$$E\{x \sim p(x) [r^*(x, \pi(x))]\} \geq E\{x \sim p(x) [r_\psi(x, \pi(x))]\} - \Delta(r_\psi, r^*)$$

Se la policy ottimizzata π supera la policy di riferimento π_{ref} di un margine Δ_{reward} sull'obiettivo r_ψ — che è la definizione operativa di miglioramento attraverso RLHF — allora la degradazione sull'obiettivo r^* è al massimo $\Delta(r_\psi, r^*)$. Questa bound è vuota se $\Delta(r_\psi, r^*) \geq \Delta_{\text{reward}}$: il guadagno osservato sul modello di ricompensa è completamente assorbito dall'errore del modello di ricompensa rispetto alla preferenza reale.

Gao, Schulman e Hilton (2022) hanno caratterizzato empiricamente la scala di questa degradazione mostrando che il reward sul modello di ricompensa cresce monotonamente con le iterazioni di RLHF mentre il reward gold (stimato da un ensemble separato di modelli di ricompensa) prima cresce e poi decresce — il fenomeno noto come *reward overoptimization*. La distanza KL ottimale che massimizza il gold reward è finita e relativamente piccola (intorno a 5–10 nat in dipendenza dal modello), il che significa che il budget di ottimizzazione oltre cui il RLHF diventa controproducente è strutturalmente limitato indipendentemente dalla qualità del modello di ricompensa.

Il framework delle etichette complementari di Yin et al. riduce $\Delta(r_\psi, r^*)$ aumentando la copertura del dataset di valutazione — più supervisor con expertise complementare coprono più regioni dello spazio $\mathcal{X} \times \mathcal{Y}$ rispetto a supervisor generici. Ma non elimina Δ : finché il dataset di supervisione è finito, la distanza Δ è positiva, e il bound di impossibilità rimane attivo. La recursive critique di Wen et al. riduce la varianza del segnale di supervisione — il modello di ricompensa addestrato su critiche di ordine superiore è più stabile — ma non riduce il bias sistematico introdotto dalle limitazioni epistemiche degli annotatori umani.

Formalmente, nessun meccanismo di supervisione che si appoggi esclusivamente su preferenze umane può garantire $\Delta(r_\psi, r^*) = 0$ in spazi di input-output non banali, a meno che il dataset di supervisione non sia esaustivo — ossia a meno che ogni coppia (x, y) non sia stata valutata da almeno un supervisore umano competente. Questa condizione non può essere soddisfatta in pratica per spazi di input-output continui o ad alta dimensionalità.

La conseguenza per il programma dell'allineamento è che le garanzie di allineamento producibili sono *approssimate*, con un errore che dipende dalla copertura del dataset di supervisione e dalla qualità degli annotatori, non *esatte* nel senso che un'analisi formale dei sistemi di controllo richiederebbe. Questo non invalida RLHF come strumento pratico — riduce l'errore in modo sistematicamente misurabile — ma fissa un confine strutturale oltre il quale le garanzie empiriche non possono spingersi.

19.6 Il confine della verifica formale per sistemi neurali di controllo

Il paper di Araiza-Illan, Eder e Richards (2014) su *Formal Verification of Control Systems' Properties with Theorem Proving* offre un punto di contrasto illuminante per comprendere la natura della lacuna tra le garanzie ottenibili con metodi formali e quelle ottenibili con metodi empirici per i sistemi JEPA.

L'approccio proposto da Araiza-Illan et al. usa lo strumento Why3 — un sistema di verifica deuttiva che interfaccia con SMT solver come CVC3 — per verificare formalmente proprietà ad alto livello di sistemi di controllo modellati in Simulink. Il meccanismo fondamentale è la traduzione del diagramma a blocchi del sistema in *teorie* nel linguaggio logico di Why3, dove ogni blocco Simulink (prodotto, somma, unità di ritardo, guadagno, comparatori) diventa un insieme di assiomi sulle relazioni input-output, e le proprietà da verificare — come la stabilità di Lyapunov — diventano *obiettivi di dimostrazione* (goal) che il theorem prover deve verificare.

Per il sistema del primo ordine discreto $x(k+1) = a \cdot x(k)$ con funzione di Lyapunov $V(x) = x^2$, la verifica della stabilità si riduce alla dimostrazione del goal:

$G1 : \forall k : \mathbb{Z}. \text{not_zero_x_op } k = \text{True} \rightarrow \text{desc_grad_op } k = \text{True}$

dove $\text{not_zero_x_op } k$ rappresenta la preconditione $x(k) \neq 0$ e $\text{desc_grad_op } k$ la postcondizione $V(x(k)) - V(x(k-1)) < 0$. Il solver CVC3 dimostra questo goal in meno di 150 secondi per valori del guadagno $a \in (-1, 1)$ — i valori per cui il sistema è stabile — e restituisce «Unknown» per $|a| \geq 1$. Il risultato è una garanzia formale: la proprietà di stabilità Lyapunov è *vera per costruzione* nello spazio logico del modello, non soltanto verificata su un insieme di simulazioni.

La distanza strutturale tra questo approccio e ciò che è applicabile ai sistemi JEPA è fondamentale. La traduzione da Simulink a Why3 funziona perché il sistema di controllo ha una struttura *dichiarativa* con componenti finiti, relazioni input-output algebriche esplicite, e proprietà definibili attraverso predicati della logica del primo ordine con aritmetica reale. Le reti neurali profonde — inclusi gli encoder e i predittori dei sistemi JEPA — non hanno nessuna di queste proprietà nella loro forma generale:

Il primo ostacolo è la *non-linearità arbitraria*: le funzioni implementate dai layer di una rete neurale profonda sono composizioni di non-linearità (ReLU, GELU, LayerNorm, Softmax) senza forma algebrica chiusa che possa essere direttamente tradotta in assiomi della logica del primo ordine. Le reti da miliardi di parametri di V-JEPA 2.1 o di OpenVLA non hanno un equivalente Simulink — sono oggetti computazionali senza struttura a blocchi dichiarativa.

Il secondo ostacolo è la *dimensionalità dello spazio latente*: le teorie di Why3 sviluppate da Araiza-Illan et al. oprano su segnali scalari o su vettori di dimensionalità bassa (tipicamente $n \leq 10$). Lo spazio latente di un encoder JEPA è tipicamente $d \in \{256, 512, 1024\}$ dimensioni, e il costo della verifica formale cresce almeno polinomialmente (spesso esponenzialmente) con la dimensionalità. La stessa verifica che richiede 150 secondi per un sistema scalare richiederebbe tempi computazionali intrattabili per sistemi ad alta dimensionalità.

Il terzo ostacolo è la *natura delle proprietà da verificare*: la stabilità di Lyapunov è una proprietà definibile in logica del primo ordine con aritmetica reale — esattamente il frammento che gli SMT solver come CVC3 e Z3 gestiscono con efficienza. Le proprietà che si vorrebbero verificare per i sistemi JEPA — causalmente disentangled, robusto agli shift di distribuzione, allineato alle preferenze dell'utente — non sono in generale esprimibili in logica del primo ordine senza approssimazioni che svuotano il loro contenuto semantico.

Il quarto ostacolo, il più fondamentale, è l'*assenza di specificazione formale*: per verificare formalmente una proprietà di un sistema JEPA, occorrerebbe prima formalizzare quella proprietà in un linguaggio logico preciso. Ma la nozione di «causalmente disentangled» non ha una definizione universalmente accettata nella comunità del Causal Representation Learning — come documentato nel Capitolo 2. Se la proprietà da verificare non ha una specificazione formale consensuale, il problema della verifica formale non è mal risolto: è mal posto.

Nonostante questi ostacoli strutturali, il lavoro di Araiza-Illan et al. identifica una direzione di ricerca che il programma JEPA non ha ancora esplorato: la verifica formale dei *moduli* del sistema che hanno struttura algebrica esplicita, come i layer di normalizzazione (LayerNorm ha una specifica algebrica precisa), le proiezioni lineari, i meccanismi di attenzione nella loro formulazione scaled dot-product. Un approccio di verifica formale *modulare* — in cui si verificano formalmente i moduli con struttura dichiarativa e si testano empiricamente i moduli non-lineari — potrebbe produrre garanzie parziali ma non triviali per i sistemi JEPA. Questa è una direzione aperta che il Capitolo 20 include nel programma di ricerca futuro.

19.7 La struttura del teorema di No Free Lunch nel dominio della rappresentazione causale: dimostrazione e analisi dell'inductive bias di JEPA

I teoremi di No Free Lunch (NFL) per l'apprendimento automatico — derivati da Wolpert e Macready (1997) nell'ambito dell'ottimizzazione e da Wolpert (1996) nell'ambito dell'apprendimento supervisionato — affermano che ogni algoritmo di apprendimento ha prestazioni medie identiche su tutti i problemi quando si mediino uniformemente su tutte le possibili funzioni da apprendere. Nel dominio del Causal Representation Learning, esiste un analogo del teorema NFL che limita le garanzie ottenibili da qualsiasi algoritmo di scoperta causale. La sezione corrente intende non soltanto enunciare questo risultato, ma costruirne la dimostrazione nella forma adatta al contesto dei sistemi JEPA, e derivarne formalmente quale inductive bias specifico di JEPA restringe il dominio in cui l'algoritmo è strutturalmente avvantaggiato.

Formalizzazione del setting. Siano $V = \{V_1, V_2, \dots, V_p\}$ le p variabili osservabili. Si denota con $\mathcal{D}_p$ l'insieme di tutti i DAG orientati aciclici su V . La cardinalità di $\mathcal{D}_p$ cresce superesponenzialmente in p : per $p = 5$, $|\mathcal{D}_p| = 29281$; per $p = 10$, $|\mathcal{D}_p| \approx 4.2 \cdot 10^{18}$. Uno Structural Causal Model (SCM) f su V è una tripla $(G_f, F_f, P_{\{\epsilon, f\}})$ dove $G_f \in \mathcal{D}_p$ è il DAG causale, $F_f = \{f_i : Pa_G(V_i) \times \epsilon_i \rightarrow V_i\}_{i=1}^p$ è la collezione delle equazioni strutturali con $Pa_G(V_i)$ i genitori di V_i nel DAG, e $P_{\{\epsilon, f\}} = \prod_{i=1}^p P_{\{\epsilon_i, f_i\}}$ è la distribuzione prodotto dei rumori esogeni, assunti mutuamente indipendenti. Si denota con $\mathcal{F}_p$ l'insieme di tutti gli SCM su p variabili in cui le equazioni strutturali sono funzioni Borel-misurabili da $\mathbb{R}^{|Pa|} \times \mathbb{R}$ a $\mathbb{R}$ e i rumori hanno densità assoluta rispetto alla misura di Lebesgue.

Un algoritmo di scoperta causale A è una funzione $A: \mathbb{R}^{n \times p} \rightarrow \mathcal{D}_p$ che mappa un dataset D di n osservazioni i.i.d. a un DAG stimato $\hat{G} = A(D)$. La qualità della stima è misurata dallo Structural Hamming Distance: $d_C(\hat{G}, G) = \#\{(i, j) : \hat{G}_{ij} \neq G_{ij}\}$ dove $\hat{G}_{ij} \in \{0, 1\}$ indica la presenza dell'arco $i \rightarrow j$ in $\hat{G}$.

Enunciato. Per qualsiasi coppia di algoritmi di scoperta causale A_1 e A_2 , per qualsiasi dataset $D \in \mathbb{R}^{n \times p}$:

$$\mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{F}_p)\}} [d_C(A_1(D), G_f)] = \mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{F}_p)\}} [d_C(A_2(D), G_f)]$$

Dimostrazione. La strategia è mostrare che, per qualsiasi algoritmo A e qualsiasi dataset D , la quantità $\mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{F}_p)\}} [d_C(A(D), G_f)]$ è una costante che non dipende né da A né da D .

Passo 1 — Decomposizione per fibra. La distribuzione uniforme $\text{Uniform}(\mathcal{F}_p)$ si fattorizza naturalmente in una distribuzione uniforme su $\mathcal{D}_p$ (la struttura del DAG) e, condizionatamente a ogni DAG G , una distribuzione uniforme sullo spazio $\mathcal{E}_{\{p, G\}}$ delle coppie $(F_f, P_{\{e, f\}})$ compatibili con G . Questa fattorizzazione segue dalla costruzione di $\mathcal{F}_p$ come prodotto cartesiano con uguale massa su ciascuna fibra. Di conseguenza:

$$\begin{aligned} \mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{F}_p)\}} [d_C(\hat{G}, G_f)] &= \sum_{\{G \in \mathcal{D}_p\}} (1/|\mathcal{D}_p|) \cdot \\ \mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{E}_{\{p, G\}})\}} [d_C(\hat{G}, G_f)] &= \sum_{\{G \in \mathcal{D}_p\}} (1/|\mathcal{D}_p|) \cdot d_C(\hat{G}, G) \end{aligned}$$

dove l'ultima eguaglianza segue dal fatto che $d_C(\hat{G}, G_f) = d_C(\hat{G}, G)$ non dipende dall'equazione strutturale f , ma soltanto dalla struttura del DAG $G_f = G$.

Passo 2 — Invarianza della somma per permutazione. Si deve ora mostrare che $S(\hat{G}) \triangleq \sum_{\{G \in \mathcal{D}_p\}} d_C(\hat{G}, G)$ è la stessa costante per qualsiasi $\hat{G} \in \mathcal{D}_p$. Sia $\hat{G}'$ un secondo elemento di $\mathcal{D}_p$. Si considera l'azione del gruppo simmetrico S_p sulle etichette dei nodi: per qualsiasi permutazione $\sigma \in S_p$, definisco $\sigma(G)$ come il DAG con arco $\sigma(i) \rightarrow \sigma(j)$ per ogni arco $i \rightarrow j$ in G . La mappa $\sigma: \mathcal{D}_p \rightarrow \mathcal{D}_p$ è una biiezione su $\mathcal{D}_p$ (permutare le etichette trasforma DAG in DAG). Per la biiezione σ si ha:

$$\begin{aligned} \sum_{\{G \in \mathcal{D}_p\}} d_C(\hat{G}, G) &= \sum_{\{G \in \mathcal{D}_p\}} d_C(\hat{G}, \sigma^{-1}(G)) \text{ [ridenominazione} \\ &\text{dell'indice di somma]} = \sum_{\{G \in \mathcal{D}_p\}} d_C(\sigma(\hat{G}), G) \text{ [poiché } d_C \text{ è invariante} \\ &\text{per ri-labeling coerente]} \end{aligned}$$

dove l'ultimo passaggio utilizza la proprietà che $d_C(\hat{G}, G') = d_C(\sigma(\hat{G}), \sigma(G'))$ per qualsiasi σ (la distanza di Hamming conta le discrepanze nelle posizioni degli archi, indipendentemente da come si nominano i nodi). Scegliendo σ tale che $\sigma(\hat{G}) = \hat{G}'$ (una tale σ esiste sempre in S_p per qualsiasi coppia di grafi sullo stesso insieme di nodi, eventualmente componendo più permutazioni), si ottiene $S(\hat{G}) = S(\hat{G}')$ per qualsiasi coppia $\hat{G}, \hat{G}'$. La somma è dunque costante: $S(\hat{G}) = C_p$ per qualsiasi $\hat{G} \in \mathcal{D}_p$, dove C_p dipende soltanto da p e dalla metrica d_C .

Passo 3 — Conclusione. Combinando i due passi:

$$\mathbb{E}_{\{f \sim \text{Uniform}(\mathcal{F}_{\mathcal{P}})\}} [d_C(A(D), G_f)] = (1/|\mathcal{D}_{\mathcal{P}}|) \cdot S(A(D)) = C_{\mathcal{P}} / |\mathcal{D}_{\mathcal{P}}|$$

La costante $C_{\mathcal{P}} / |\mathcal{D}_{\mathcal{P}}|$ è indipendente da A e da D . Poiché vale per qualsiasi A_1 e A_2 , segue che:

$$\mathbb{E}[d_C(A_1(D), G_f)] = \mathbb{E}[d_C(A_2(D), G_f)] = C_{\mathcal{P}} / |\mathcal{D}_{\mathcal{P}}| \quad \square$$

Osservazione sulle ipotesi. La dimostrazione presuppone che $\mathcal{F}_{\mathcal{P}}$ sia invariante per permutazione delle variabili: ogni struttura causale è considerata plausibile indipendentemente dal labeling dei nodi. Questa ipotesi di simmetria è essenziale e non è sempre soddisfatta in contesti applicativi dove la semantica delle variabili vincola le strutture plausibili. Il NFL si applica pertanto alla situazione di ignoranza a priori completa sulla struttura del dominio — esattamente la condizione che caratterizza l'applicazione di JEPA a domini nuovi senza conoscenza preventiva.

L'inductive bias specifico di JEPA e la caratterizzazione formale di $\mathcal{F}^+$. Il NFL causale stabilisce che nessun algoritmo è mediamente superiore agli altri su $\mathcal{F}_{\mathcal{P}}$ nella sua interezza. Ogni algoritmo con inductive bias specifico — e JEPA ne possiede tre distinti e formalizzabili — è però superiore alla media su un sottoinsieme $\mathcal{F}^+ \subseteq \mathcal{F}_{\mathcal{P}}$ e inferiore alla media su $\mathcal{F}_{\mathcal{P}} \setminus \mathcal{F}^+$. La caratterizzazione formale di $\mathcal{F}^+$ richiede di rendere esplicite le proprietà degli SCM che l'inductive bias di JEPA implicitamente presuppone.

Il primo inductive bias riguarda la *continuità delle transizioni*. L'obiettivo $\mathcal{L}_{\text{JEPA}} = \mathbb{E}_t [\|\hat{z}_{t+1} - z_{t+1}\|_2^2]$ è minimizzato quando il predittore $\text{pred}_{\varphi}(z_t, a_t)$ approssima z_{t+1} accuratamente, il che è possibile soltanto se le transizioni nello spazio latente sono approssimabili da una funzione liscia ($z_t, a_t \mapsto z_{t+1}$). Formalmente, il predittore converge a una funzione $\text{pred}_{\varphi^*} \in L_2$ soltanto se le transizioni del processo generativo soddisfano una condizione di Lipschitz con costante $L_{\text{trans}} < \infty$:

$$\|z_{t+1} - z_{t'+1}\| \leq L_{\text{trans}} \cdot (\|z_t - z_{t'}\| + \|a_t - a_{t'}\|)$$

Gli SCM le cui equazioni strutturali producono transizioni non-Lipschitz — con discontinuità, effetti soglia, comportamento impulsivo — si trovano al di fuori di $\mathcal{F}^+$ perché il predittore di JEPA non può approssimarli senza errore sistematico elevato. Anche se la struttura del DAG fosse identificabile in linea di principio, l'errore di predizione per queste transizioni invalida la stima causale.

Il secondo inductive bias riguarda la *gaussianità isotropa dei fattori latenti*, imposta da SIGReg. Questo termine è derivato dalla condizione di identificabilità dell'iVAE (Khemakhem et al. 2020): l'ottimizzazione di $\text{Reg}_{\text{SIG}} = \mathbb{E}_{\{j \in [d]\}} [W^2(\mathcal{N}_j, \mathcal{M}(0,1))]$ equivale a imporre che la distribuzione marginale degli embedding soddisfi la condizione di normalità isotropa che rende il problema di identificabilità ben posto. Gli SCM con distribuzioni di rumore pesante-coda (Cauchy, Pareto), bimodali, o discreti si trovano al di fuori di $\mathcal{F}^+$ perché

l'encoder minimizzerà Reg_SIG deformando la struttura dei fattori latenti verso la gaussianità impostale, anziché rispecchiare la vera struttura causale. La condizione iVAE richiede inoltre che i fattori esponenziali che parametrizzano il rumore varino con la variabile ausiliaria (segmento temporale o contesto): SCM senza questa variabilità contestuale strutturata non soddisfano il criterio di identificabilità.

Il terzo inductive bias riguarda la *esogeneità delle azioni*. Il DAG JEPA non include archi da z_t ad a_t , presupponendo che le azioni siano esogene rispetto allo stato latente — ovvero che la distribuzione delle azioni nel dataset di training non dipenda da variabili confondenti nascoste non catturate da z_t . Gli SCM in cui le azioni sono endogene — scelte da una politica $\pi(a_t | S_t)$ che dipende da componenti di S_t non osservate e non catturate da $z_t = \text{enc}_\theta(o_t)$ — producono una distribuzione di training in cui $p(z_{t+1} | z_t, a_t)$ confonde effetti causali e correlazioni spurie dovute alla politica di raccolta dati. In questo regime il predittore apprende la distribuzione osservazionale condizionata, non la distribuzione interventale, vanificando la connessione con il do-calcolo stabilita nel Capitolo 17.

La struttura formale di $\mathcal{F}^*$ e la sua complementare. Le tre condizioni precedenti definiscono:

$\mathcal{F}^* = \{f \in \mathcal{F}_p : \text{equazioni strutturali } L_{\text{trans-Lipschitz}} \wedge \text{distribuzioni di rumore soddisfano le condizioni iVAE di Khemakhem et al.} \wedge \text{azioni esogene nel DAG di } f\}$

Il NFL causale implica la seguente relazione di compensazione: se A_{JEPA} supera il livello aleatorio su $\mathcal{F}^*$ — ovvero se

$$\mathbb{E}_{f \sim \text{Uniform}(\mathcal{F})} [d_C(A_{\text{JEPA}}(D), G_f)] < C_p / |\mathcal{D}_p|$$

— allora necessariamente:

$$\mathbb{E}_{f \sim \text{Uniform}(\mathcal{F}_p \setminus \mathcal{F}^*)} [d_C(A_{\text{JEPA}}(D), G_f)] > C_p / |\mathcal{D}_p|$$

Il vantaggio di JEPA su $\mathcal{F}^*$ è bilanciato esattamente da un deficit su $\mathcal{F}_p \setminus \mathcal{F}^*$ tale che la media pesata sulla distribuzione uniforme rimanga la costante $C_p / |\mathcal{D}_p|$. Le proprietà degli spazi JEPA dimostrate su benchmark come 3DIEBench — dove la struttura causale rispetta per costruzione l'inductive bias (fattori indipendenti, variazioni strutturate e continue, trasformazioni di gruppo) — non si trasferiscono quindi a strutture causali che violano anche soltanto una delle condizioni di $\mathcal{F}^*$. Senza una verifica preliminare dell'appartenenza a $\mathcal{F}^*$ — test di Lipschitz-continuità delle transizioni osservate, test di gaussianità delle distribuzioni marginali dei fattori latenti, analisi della politica di raccolta dati per verificare l'assenza di confondenti nascosti — è formalmente impossibile determinare ex ante se l'algoritmo JEPA si trovi nel regime favorevole o in quello sfavorevole della partizione imposta dal NFL.

19.8 Il problema del collasso della distribuzione di preferenze e la degenerazione del reward model

Un limite strutturale del paradigma RLHF che si connette direttamente con le proprietà degli spazi JEPA riguarda la degenerazione del modello di ricompensa sotto ottimizzazione. Il fenomeno è noto come *overoptimization* del reward model, e Gao, Schulman e Hilton (2022) ne hanno caratterizzato la dipendenza dalla dimensionalità del modello: la scala del reward ottimale prima che l'overoptimization diventi rilevante dipende dal numero di parametri del modello di ricompensa, ma in modo sublineare — raddoppiare i parametri non raddoppia la robustezza all'overoptimization.

Per i sistemi JEPA specificamente, il problema si manifesta nella tensione tra la struttura di equivarianza dell'encoder e la distribuzione delle preferenze umane. Come argomentato nel Capitolo 15, un valutatore umano esprime preferenze che possono essere sistematicamente correlate con caratteristiche superficiali delle rappresentazioni visive — la prospettiva da cui è mostrata la scena, la luminosità dell'immagine, la presenza di distrattori visivi — che sono variabili trasformazionali filtrate dalla struttura di invarianza di JEPA. Un modello di ricompensa addestrato su queste preferenze «impara» a privilegiare caratteristiche superficiali che gli annotatori preferiscono inconsciamente, producendo segnali di reward che sono sistematicamente distorti rispetto alle proprietà causalmente rilevanti del task.

La degenerazione specifica che può emergere in un sistema JEPA con RLHF può essere formalizzata attraverso la divergenza tra la distribuzione delle preferenze umane e la metrica sullo spazio latente JEPA. Sia $d_{\text{JEPA}}(z, z')$ la distanza coseno sullo spazio latente JEPA — la metrica che l'encoder ha appreso durante il preaddestramento e che codifica le invarianze semanticamente rilevanti. Sia $d_{\text{pref}}(y, y')$ la distanza sullo spazio degli output misurata attraverso le preferenze umane — la metrica implicita che il modello di ricompensa r_{ψ} apprende dal dataset di confronti. Se queste due metriche non sono concordanti — se $d_{\text{JEPA}}(z_1, z_2) < d_{\text{JEPA}}(z_1, z_3)$ (z_2 è semanticamente più simile a z_1 di z_3) mentre $d_{\text{pref}}(y_1, y_2) > d_{\text{pref}}(y_1, y_3)$ (gli annotatori preferiscono y_3 a y_2 rispetto a y_1) — allora il fine-tuning con RLHF spinge il sistema verso output che violano la struttura di invarianza appresa durante il preaddestramento.

La condizione formale perché RLHF non degradi la struttura JEPA è che la metrica delle preferenze umane sia *compatibile* con la metrica sullo spazio latente JEPA — che le preferenze umane rispettino le invarianze che JEPA ha appreso. Questa compatibilità non può essere imposta dall'architettura: dipende dalla struttura delle preferenze umane effettive degli annotatori, che sono prodotte da processi cognitivi e culturali non controllabili dal progettista del sistema. Il Capitolo 15 aveva proposto la raccolta di *preference data che rispettano l'equivarianza*

come strategia parziale, ma questa strategia richiede una comprensione preventiva di quale struttura di invarianza il sistema JEPA ha appreso — una comprensione che richiede probing causale estensivo prima della raccolta dei dati di preferenza. Questo circolo — si deve capire la struttura per raccogliere i dati giusti, ma si capisce la struttura soltanto dopo averla analizzata — è una forma di bootstrap epistemico che non ha una soluzione semplice.

19.9 La limitazione della scalable oversight attraverso la composizione di segnali deboli

Un limite strutturale che emerge dalla combinazione dei due framework di scalable oversight esaminati — il framework delle etichette complementari di Yin et al. e il framework della recursive critique di Wen et al. — riguarda la *composizione* di segnali di supervisione deboli. Entrambi i framework trasformano segnali di supervisione deboli (l'etichetta complementare che esclude un'opzione, la critica di ordine superiore che valuta una critica) in stime dell'accuratezza del sistema. La domanda è: questi segnali si compongono in modo tale che la loro combinazione fornisca garanzie più forti di ciascuno individualmente?

La risposta è negativa in generale, per una ragione che riguarda la struttura delle correlazioni tra i segnali. Le etichette complementari fornite da esperti di domini diversi non sono indipendenti condizionatamente all'input: due esperti di domini vicini (per esempio cardiologia e pneumologia in un task di diagnosi interna) tendono a produrre esclusioni simili perché le loro competenze si sovrappongono parzialmente. Questa correlazione viola l'assunzione di uniformità della distribuzione delle etichette complementari — non nell'assegnazione dell'indice (che è randomizzata dal protocollo), ma nell'accuratezza con cui l'etichetta riflette la vera limitazione epistemica dell'esperto. Formalmente: se $\text{Cov}(W_i, W_j) \neq 0$ per annotatori i, j con expertise correlata, la varianza dello stimatore aggregato non è ridotta della quantità prevista dalla formula di Yin et al. per annotatori indipendenti, ma è limitata inferiormente da:

$$\text{Var}(\hat{A}_{\text{comp}}) \geq (K-1)^2 \cdot [q(1-q)/n_c + (1/n_c^2) \cdot \sum_{\{i \neq j\}} \text{Cov}(W_i, W_j)]$$

Il termine di covarianza è positivo quando le competenze degli annotatori sono correlate, il che significa che la varianza effettiva dello stimatore è maggiore di quella prevista dalla teoria con annotatori indipendenti. Se le covarianze sono sistematicamente grandi — come ci si attende in domini dove l'expertise è altamente specializzata e settorializzata — il vantaggio quantitativo delle etichette complementari rispetto alle ordinarie si riduce significativamente rispetto alle previsioni teoriche.

Le garanzie finite-sample di Theorem 2 di Yin et al. sono valide sotto l'assunzione di campionamento i.i.d. delle variabili W_i , che è garantita dal

protocollo di selezione uniforme degli annotatori. Ma la qualità semantica delle etichette — la loro capacità di riflettere genuina conoscenza di esclusione del dominio — non è garantita dal protocollo, e può variare sistematicamente con caratteristiche degli annotatori non misurate nel design sperimentale. Questa variabilità non-campionistica introduce un bias nella stima dell'accuratezza che i bound di deviazione di Yin et al. non catturano — una lacuna strutturale della teoria che diventa rilevante nella pratica quando gli annotatori sono umani con limitazioni cognitive e pregiudizi impliciti.

19.10 Il confine del paradigma: quando i limiti strutturali si sovrappongono

I limiti documentati nelle sezioni precedenti non sono indipendenti: formano una struttura di interdipendenze che definisce il confine del paradigma JEPA come fondamento per l'intelligenza incarnata e verificabile. Questa struttura merita un'analisi sintetica prima di passare al programma di ricerca futuro.

Il limite dell'identificabilità causale (Sezione 19.3) impone che i fattori latenti apprendibili siano limitati dall'informazione portata dalle variabili ausiliarie disponibili — le azioni, nel caso di LeWorldModel e seq-JEPA. Il limite dell'allineamento senza supervisione esaustiva (Sezione 19.5) impone che le garanzie di conformità ai valori siano limitate dalla copertura del dataset di supervisione. Il limite della verifica formale (Sezione 19.6) impone che le garanzie formali siano limitate alla verificabilità delle proprietà strutturalmente semplici. Il limite del No Free Lunch causale (Sezione 19.7) impone che nessun inductive bias sia universalmente superiore agli altri. Il limite della compatibilità metriche preferenze-spazio latente (Sezione 19.8) impone che RLHF possa degradare le proprietà strutturali apprese durante il preaddestramento JEPA.

La sovrapposizione di questi limiti produce un corollario che il programma di ricerca sull'intelligenza incarnata e verificabile deve riconoscere esplicitamente: non esiste, in linea di principio, un sistema che sia simultaneamente identificabile causalmente senza assunzioni su tutto il dominio, allineato alle preferenze umane senza supervisione esaustiva, verificabile formalmente per proprietà arbitrarie, universalmente robusto attraverso inductive bias, e immune dalla degradazione del reward model. Questi obiettivi sono in tensione strutturale l'uno con l'altro — non contingentemente ma per ragioni che derivano dai fondamenti matematici dei problemi che affrontano.

Riconoscere questa tensione non significa abbandonare il programma — significa costruirlo su basi epistemologicamente oneste. Il programma di ricerca futuro (Capitolo 20) che emerge da questo inventario dei limiti non punta all'eliminazione delle tensioni ma alla loro gestione principiata: identificare le condizioni sotto cui i diversi obiettivi possono essere soddisfatti simultaneamente in modo approssimato, quantificare i compromessi tra gli

obiettivi, e costruire sistemi che siano trasparenti riguardo a quali garanzie forniscono e su quale porzione del loro dominio operativo.

Capitolo 20 — Il programma di ricerca futuro: dalle lacune dimostrate alle priorità sperimentali

20.1 La struttura di un programma di ricerca emergente dai propri limiti

Il modo più onesto di formulare un programma di ricerca futuro non è elencare le direzioni promettenti in senso astratto ma derivarle direttamente dalle lacune dimostrate nell'analisi precedente. Il Capitolo 19 ha inventariato i limiti strutturali del paradigma JEPA come fondamento per l'intelligenza incarnata e verificabile: l'identificabilità parziale dei fattori latenti, il confine della supervisione umana, l'inaccessibilità della verifica formale per sistemi neurali complessi, l'impossibilità di composizione senza perdita di segnale debole. Il presente capitolo trasforma ciascuno di questi limiti in un sotto-programma di ricerca con obiettivi sperimentali verificabili, metodi proposti, e connessioni con gli strumenti teorici già sviluppati nelle parti precedenti della monografia.

La struttura del programma è determinata dalla gerarchia logica dei problemi. Alcuni sotto-programmi sono *prerequisiti* di altri: il testing sistematico dell'ipotesi di trasferimento fisico di V-JEPA su task di controllo in simulazione è prerequisito della valutazione del sim-to-real gap in ambienti fisici reali; la verifica empirica della previsione del confine $B = C$ per la recursive critique è prerequisito della progettazione di sistemi di oversight che combinino etichette complementari e critica ricorsiva. Altri sotto-programmi sono invece *complementari* nel senso che producono evidenza su assi ortogonali dello stesso problema e si rafforzano reciprocamente senza dipendenze logiche dirette.

Il programma non pretende di essere esaustivo: la letteratura del settore si rinnova su scale temporali di mesi, e qualsiasi elenco di priorità è destinato ad essere parzialmente obsoleto entro il termine della stesura. L'obiettivo è invece indicare le linee di forza che emergono dalla struttura teorica costruita in questa monografia — le direzioni che sono più naturalmente in continuità con il programma analizzato e che, per ragioni derivabili dall'analisi, hanno le migliori probabilità di produrre evidenza genuinamente nuova.

20.2 Sotto-programma I: CV-JEPA come architettura sperimentale per l'identificabilità causale

Il framework CV-JEPA derivato nel Capitolo 3 ha una struttura teorica precisa — l'ELBO causale a sette termini, il vincolo di aciclicità NOTEARS, la regolarizzazione causale — ma non è stato ancora implementato e valutato su nessun dataset. Questa è la lacuna più direttamente colmabile con gli strumenti già disponibili: l'implementazione richiede la combinazione di componenti open-source esistenti (l'architettura JEPA di Assran et al., il constraint NOTEARS di Zheng et al., l'IVAE di Khemakhem et al.) senza sviluppi teorici ulteriori.

Il design sperimentale prioritario per CV-JEPA è una valutazione su 3DIEBench nella variante interventoriale, confrontando CV-JEPA con: CausalVAE standard (Yang et al., 2021) come baseline di riferimento per l'identificabilità causale impostata esplicitamente; seq-JEPA come baseline JEPA senza la componente causale esplicita; iVAE standard come baseline di identificabilità causale senza la componente predittiva JEPA. Le metriche principali sono tre: l' R^2 di equivarianza per la previsione dei fattori causali dai fattori latenti, l'accuratezza del classificatore lineare sulla categoria degli oggetti come misura di invarianza, e il coefficiente di Spearman tra le distribuzioni interventoriali computate dal modello e quelle osservate nel dataset.

Il training objective di CV-JEPA si scrive nella forma compatta:

$$\mathcal{L}_{\text{CV-JEPA}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{regol}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{causal}}$$

dove i termini sono definiti nel Capitolo 3 (equazione 3.3). La componente computazionalmente più costosa è $\mathcal{L}_{\text{causal}} = -\gamma \cdot h(A_c, A_t)$ con:

$$h(A_c, A_t) = \text{tr}(\exp(A_c \circ A_c)) + \text{tr}(\exp(A_t \circ A_t)) - 2n$$

dove $A_c, A_t \in \mathbb{R}^{n \times n}$ sono le matrici di adiacenza dei DAG per le viste contesto e target, e $\circ$ denota il prodotto di Hadamard. Il calcolo dell'esponenziale matriciale richiede $O(n^3)$ operazioni, con n tipicamente nella fascia 32–128 per problemi di scala ragionevole. Il costo è quindi trattabile per dataset come 3DIEBench ma richiederebbe ottimizzazioni strutturali (approssimazioni sparse, tecniche di block-diagonal) per scalare a encoder ad alta dimensionalità come quelli di V-JEPA 2.1.

La previsione teorica derivabile dall'analisi del Capitolo 17 è che CV-JEPA migliori rispetto a seq-JEPA sull' R^2 di equivarianza nelle regioni dello spazio degli stati dove la struttura del DAG è più vincolante — ossia dove la dipendenza causale tra i fattori è più forte. Per i fattori causalmente indipendenti (colore del pavimento e rotazione dell'oggetto in 3DIEBench, che sono generati indipendentemente), CV-JEPA non dovrebbe mostrare vantaggi rispetto a seq-JEPA perché il vincolo di aciclicità non aggiunge informazione su strutture già indipendenti. Il vantaggio dovrebbe concentrarsi per fattori con dipendenze causali non banali — per esempio in scenari dove la posizione di un oggetto dipende causalmente dalla sua velocità nel frame precedente — che sono il caso di test naturale per il programma futuro.

20.3 Sotto-programma II: Sparse Autoencoders sugli spazi JEPA visivi

Il Capitolo 14 ha derivato le previsioni teoriche per il comportamento degli SAE sugli spazi latenti JEPA biforcati (z_i equivariante e z_{AGG} invariante), ma ha constatato che non esiste ancora una letteratura sistematica su SAE applicati a encoder visivi JEPA — la letteratura sugli SAE si concentra sui LLM. Costruire questa letteratura è una priorità del programma futuro per due ragioni: la prima è che le previsioni teoriche del Capitolo 14 sono verificabili con metodi già sviluppati per i LLM (adattati al dominio visivo); la seconda è che i risultati empirici potrebbero differire significativamente dalle previsioni teoriche, rivelando proprietà dello spazio JEPA non catturate dall'analisi teorica.

Il design sperimentale proposto procede per fasi. La prima fase addestra un SAE standard — senza la modifica equivariante proposta nel Capitolo 14 — sulle attivazioni del layer finale di un encoder seq-JEPA pre-addestrato su 3DIEBench. Il SAE ha $n \gg d$ feature (tipicamente $n = 4d$ o $n = 8d$), funzione di attivazione ReLU, e training objective:

$$\mathcal{L}_{SAE}(z) = \|z - \text{dec}(\text{enc}(z))\|^2 + \alpha \cdot \|\text{enc}(z)\|_1$$

con $\alpha \in \{0.001, 0.003, 0.01, 0.03\}$ come valori di ricerca degli iperparametri. La metrica principale è il monosemanticity score per le feature attivate sui top-K input del dataset di test: si visualizzano i $K = 10$ patch che attivano maggiormente ciascuna feature, si verifica se condividono una proprietà visiva comune (una norma di superficie, un gradiente di colore, un bordo orientato), e si calcola la proporzione di feature monosemantiche come:

$$\text{mono_score}(\text{SAE}) = |\{k : \text{monosemantic}(k)\}| / n$$

dove $\text{monosemantic}(k)$ è una valutazione umana binaria sull'interpretabilità della feature k dal suo set di top-K attivazioni. La previsione del Capitolo 14 è che mono_score sia più alto per lo spazio z_{AGG} che per lo spazio z_i , e che le feature monosemantiche di z_i corrispondano a proprietà locali (texture, bordi) mentre quelle di z_{AGG} corrispondano a proprietà semantiche (categoria, forma globale).

La seconda fase addestra il SAE equivariante proposto nel Capitolo 14, con il termine aggiuntivo di regolarizzazione sulle orbite di trasformazione:

$$\mathcal{L}_{SAE_equiv}(z) = \|z - \text{dec}(\text{enc}(z))\|^2 + \alpha \cdot \|\text{enc}(z)\|_1 + \beta \cdot \sum_k \|\text{enc}_k(z) - \text{enc}\{g(k)\}(g \cdot z)\|^2$$

e verifica se le feature equivarianti così addestrate corrispondono meglio alle orbite delle trasformazioni rispetto alle feature del SAE standard, misurato attraverso l'orbit consistency score:

$$\text{orbit_score}(k) = \|\text{Cov}(\hat{h}_k(z), \hat{h}_{\{g(k)\}}(g \cdot z))\|_F / (\|\text{Var}(\hat{h}_k)\| \cdot \|\text{Var}(\hat{h}_{\{g(k)\}})\|)$$

dove Cov è la covarianza tra l'attivazione della feature k sull'input z e l'attivazione della feature $g(k)$ sull'input trasformato $g \cdot z$, e la norma di Frobenius normalizza la scala.

La connessione con il programma di interpretabilità meccanicistica dei LLM — in particolare con il lavoro di Templeton et al. (2024) sugli SAE applicati a Claude — rende questo sotto-programma particolarmente rilevante per la dimensione dell'allineamento: se le feature SAE degli spazi JEPA visivi mostrano monosemanticità comparabile a quelle trovate per i LLM, si apre la possibilità di costruire interpretazioni meccanicistiche complete dei sistemi JEPA — comprendere non soltanto cosa rappresentano le rappresentazioni (probing) ma come e perché vengono compute (meccanismo).

20.4 Sotto-programma III: il test sistematico dell'ipotesi di trasferimento fisico di V-JEPA

Il Capitolo 8 ha enunciato l'*ipotesi di trasferimento fisico di JEPA*: un encoder V-JEPA preaddestrato su video reali di interazioni fisiche umane, quando utilizzato come Perception Module in un sistema addestrato in simulazione, dovrebbe ridurre il visual gap e il dynamics gap rispetto a un encoder addestrato solo in simulazione, perché porta dal preaddestramento una rappresentazione delle dinamiche fisiche reali che non dipende dalla fedeltà del rendering simulato. Questa ipotesi non è ancora stata testata sistematicamente.

Il design sperimentale proposto per testare questa ipotesi è un esperimento a due condizioni, idealmente condotto su un task di manipolazione con un robot Franka Panda in un ambiente simulato (MuJoCo o IsaacSim) valutato nel mondo fisico reale. Nella condizione A si usa un encoder DINOv2 preaddestrato su ImageNet (baseline), nella condizione B si usa un encoder V-JEPA 2.1 preaddestrato su dati video fisici reali (Ego4D, Kinetics). In entrambe le condizioni il World Model latente e la policy vengono addestrati su traiettorie simulate. La valutazione confronta il success rate sui task di manipolazione reale dopo lo stesso numero di step di training in simulazione.

Le metriche secondarie misurano la struttura del sim-to-real gap nelle rappresentazioni: per una sequenza di frame corrispondenti (stessa scena fisica, acquisita sia in simulazione sia nel mondo reale), si calcola la divergenza di Jensen-Shannon tra le distribuzioni delle rappresentazioni latenti nelle due condizioni:

$$\text{JS_gap}(E, \text{condizione}) = \frac{1}{2} \cdot D_{\text{KL}}(P_{\text{sim}}(E(x)) \parallel M) + \frac{1}{2} \cdot D_{\text{KL}}(P_{\text{real}}(E(x)) \parallel M)$$

con $M = \frac{1}{2}(P_{\text{sim}} + P_{\text{real}})$. La previsione dell'ipotesi di trasferimento fisico è che JS_gap sia significativamente più basso per l'encoder V-JEPA 2.1 rispetto a DINOv2, a parità di task e architettura del World Model.

Un controllo critico dell'esperimento è la verifica che la riduzione del JS_gap si traduca effettivamente in improvement sul task di manipolazione reale e non sia un artefatto della metrica di distribuzione. Se il JS_gap si riduce ma il success rate non migliora, significa che la struttura del gap che V-JEPA 2.1 riduce non è la componente del gap causalmente rilevante per il task — un risultato comunque informativo per comprendere quale tipo di discrepanza domini effettivamente le prestazioni real-world.

La connessione con l'analisi del sim-to-real gap del Capitolo 8 permette di segmentare la riduzione del gap nelle sue componenti: visual gap (misurabile attraverso la qualità della segmentazione semantica su frame reali con il modello addestrato in simulazione), dynamics gap (misurabile attraverso l'errore di previsione del World Model su traiettorie reali vs. simulate), e latency gap (misurabile attraverso il degradation profile al variare della frequenza di inferenza). La decomposizione del contributo di V-JEPA 2.1 alle singole componenti — possibile con questo design sperimentale controllato — produce una caratterizzazione più fine dell'ipotesi di trasferimento fisico rispetto a una valutazione aggregata del success rate.

20.5 Sotto-programma IV: verifica formale modulare per sistemi JEPA attraverso il paradigma Why3

Il Capitolo 19 ha identificato la verifica formale come un confine strutturale per i sistemi neurali profondi: le proprietà che possono essere verificate formalmente con strumenti come Why3 sono limitate ai moduli con struttura algebrica dichiarativa. Ma questa limitazione non significa che la verifica formale sia irrilevante per il programma JEPA — significa che deve essere applicata selettivamente ai moduli verificabili, costruendo garanzie parziali ma non triviali.

Il programma di verifica formale modulare proposto ha tre livelli. Al primo livello si verificano formalmente le componenti con struttura algebrica esplicita: la proiezione di positional encoding (che ha la forma $pe(i, 2k) = \sin(i/10000^{\{2k/d\}})$ e $pe(i, 2k+1) = \cos(i/10000^{\{2k/d\}})$, esprimibile come composizione di funzioni trascendenti elementari), il meccanismo di attenzione scaled dot-product nella sua versione deterministica (dove il softmax è sostituito da un'operazione di normalizzazione algebrica approssimata), e le operazioni di Layer Normalization (che hanno la forma $ln(x) = (x - \mu) / \sigma \cdot \gamma + \beta$ con μ e σ calcolati sull'input, verificabile in logica del primo ordine con aritmetica reale).

Per questi moduli, l'approccio di Araiza-Illan et al. è direttamente applicabile: si specifica il comportamento atteso come teoria Why3, si genera automaticamente il goal di verifica, e si invoca un SMT solver per verificarne la correttezza. Le proprietà verificabili includono la stabilità numerica (il denominatore in LayerNorm è sempre positivo per qualsiasi input reale?) e i limiti sull'output (la distanza coseno tra vettori normalizzati è sempre in $[-1, 1]$). Queste sembrano proprietà banali, ma la loro verifica formale è importante per garantire l'assenza di overflow numerici, divisioni per zero e altre condizioni di errore che le reti neurali profonde possono incontrare in inferenza su input fuori dalla distribuzione di training.

Il secondo livello estende la verifica formale alle proprietà dei predittori lineari che compongono le teste di attenzione: ogni testa di attenzione calcola $\text{Att}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$, che è una composizione di operazioni lineari e un softmax. Le proprietà lineari dei moduli di query, key e value — $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, $W_V \in \mathbb{R}^{d \times d_v}$ — sono verificabili attraverso il do-calcolo: si può dimostrare formalmente che se l'input z_t è perturbato di un vettore δ , il cambio nell'output di attenzione è linearmente approssimato da $\Delta_{\text{Att}} \approx J_{\text{Att}} \cdot \delta$ dove J_{Att} è la jacobiana del meccanismo di attenzione calcolata al punto z_t . Questa linearizzazione locale è verificabile formalmente nell'intorno di un punto specifico, anche se non globalmente.

Il terzo livello, il più ambizioso, è la verifica formale della proprietà di equivarianza dei sistemi seq-JEPA. La proprietà di equivarianza dell'encoder f_θ rispetto a un gruppo di trasformazioni G afferma che:

$$\forall g \in G, \forall x \in \mathcal{X}: f_\theta(g \cdot x) = \varrho_g \cdot f_\theta(x)$$

dove ϱ_g è la rappresentazione del gruppo G nello spazio latente. Per G finito e ϱ_g lineare, questa proprietà è esprimibile in logica del primo ordine e potenzialmente verificabile con SMT solver — non sull'intera rete neurale (che sarebbe intrattabile), ma su approssimazioni lineari locali dell'encoder in punti specifici del dominio di input. Questo approccio produce garanzie *locali* di equivarianza — proprietà che valgono in intorni di punti specifici ma non globalmente — che sono comunque più forti delle sole garanzie empiriche di probing.

La connessione con il lavoro di Araiza-Illan et al. è diretta: il loro strumento di traduzione automatica da Simulink a Why3 potrebbe essere esteso per tradurre le componenti strutturalmente semplici di un sistema JEPA (le operazioni di positional encoding, normalizzazione, e proiezione lineare) in teorie Why3, produrre automaticamente i goal di verifica, e invocare il solver. L'estensione richiede la costruzione di una libreria di teorie Why3 per le operazioni neurali elementari — analoga alla libreria di teorie per i blocchi Simulink sviluppata da Araiza-Illan et al. — che è un contributo ingegneristico fattibile con gli strumenti esistenti.

20.6 Sotto-programma V: integrazione di recursive critique ed etichette complementari per il scalable oversight ibrido

Il Capitolo 16 ha analizzato il framework delle etichette complementari di Yin et al. e il Capitolo 19 ha analizzato il framework della recursive critique di Wen et al. come approcci distinti al scalable oversight. Un sotto-programma di ricerca naturale è la loro integrazione in un sistema ibrido che sfrutti i vantaggi di entrambi: la copertura statistica delle etichette complementari (uno stimatore non distorto dell'accuratezza senza ground truth) e la profondità epistemica della recursive critique (un meccanismo di supervisione che diventa più efficace al crescere del gap di capacità tra sistema e supervisore nel regime weak-to-strong).

L'ipotesi alla base di questa integrazione è che le due modalità di supervisione siano complementari nel senso tecnico preciso: le etichette complementari producono stime affidabili dell'accuratezza aggregata ma non forniscono indicazioni su *quali* aspetti specifici dell'output del sistema siano errati; la recursive critique fornisce analisi strutturate degli errori ma produce stime rumorose dell'accuratezza complessiva (il Capitolo 19 ha documentato che lo stimatore basato su critiche ha varianza più alta di quello basato su etichette complementari quando il modello di valutazione è debole rispetto al sistema valutato). La combinazione dovrebbe ridurre sia la varianza della stima (attraverso le etichette complementari) sia il bias sistematico dell'analisi degli errori (attraverso la recursive critique).

Il design sperimentale proposto è il seguente. Per ciascun item di valutazione, si raccolgono: K etichette complementari da K esperti di diversi sotto-domini; una critica di primo ordine C^1 da un valutatore generale; una critica di secondo ordine C^2 dalla valutazione della C^1 da parte di un secondo valutatore. La stima dell'accuratezza ibrida è:

$$\hat{A}_{\text{hybrid}} = w_{\text{comp}} \cdot \hat{A}_{\text{comp}} + w_{\text{crit}} \cdot \hat{A}_{\text{crit}} + (1 - w_{\text{comp}} - w_{\text{crit}}) \cdot \hat{A}_{\text{mix}}$$

dove $\hat{A}_{\text{comp}}$ è lo stimatore complementare di Yin et al., $\hat{A}_{\text{crit}}$ è la proporzione di output preferiti nelle valutazioni della C^1 , e $\hat{A}_{\text{mix}}$ è lo stimatore di massima verosimiglianza congiunta. I pesi w_{comp} e w_{crit} sono stimati attraverso cross-validazione su un piccolo dataset con ground truth noto — una delle poche informazioni che si assumono disponibili nel regime di scalable oversight.

La previsione teorica per questo sistema ibrido è che $\text{Var}(\hat{A}_{\text{hybrid}}) < \text{Var}(\hat{A}_{\text{comp}}) < \text{Var}(\hat{A}_{\text{crit}})$, ossia che il sistema ibrido abbia varianza inferiore a ciascuno dei suoi componenti presi separatamente, per la ragione standard che combinazioni di stimatori indipendenti e non distorti hanno varianza inferiore a

ciascun componente. La condizione di indipendenza è approssimata ma non esatta — i valutatori che producono le etichette complementari e quelli che producono le critiche possono condividere pregiudizi sistematici — e la quantificazione di questa correlazione è uno degli obiettivi empirici primari del sotto-programma.

Il test su task reali di supervisione in domini specialistici — seguendo il design di Yin et al. con EDINET-Bench e Medical Abstracts — permetterebbe di valutare se l'approccio ibrido produce stime più stabili dell'accuratezza del sistema valutato rispetto ai due approcci presi separatamente. Se il miglioramento si materializza, il sistema ibrido potrebbe costituire il candidato più naturale per il framework di scalable oversight applicato a sistemi AMI operativi in domini ad alta specializzazione.

20.7 Sotto-programma VI: seq-JEPA come Perception Module nei sistemi VLA di terza generazione

Il Capitolo 5 ha formulato l'ipotesi che un encoder V-JEPA — o una combinazione V-JEPA + SigLIP — possa servire come Perception Module in un sistema VLA di terza generazione, fondendo comprensione dinamica del movimento con comprensione semantica degli oggetti. Questa ipotesi è teoricamente motivata ma non ancora sistematicamente valutata. Un sotto-programma dedicato alla sua valutazione empirica è il passo più diretto verso la connessione tra il paradigma JEPA e le architetture VLA.

Il design sperimentale proposto parte dall'architettura OpenVLA come baseline — con il suo encoder SigLIP + DINOv2 — e la confronta con una variante che sostituisce DINOv2 con un encoder seq-JEPA preaddestrato su sequenze video egocentriche di interazioni fisiche. La sostituzione è motivata dal fatto che DINOv2 e seq-JEPA condividono l'obiettivo di apprendere feature spaziali di basso livello in modo autosupervisionato, ma seq-JEPA ha una proprietà aggiuntiva: le sue rappresentazioni codificano informazione sulle trasformazioni relative tra frame successivi — le azioni, nella terminologia di seq-JEPA — che sono direttamente rilevanti per il controllo robotico.

La previsione è che la sostituzione di DINOv2 con seq-JEPA produca guadagni nei task che richiedono ragionamento sulle traiettorie fisiche degli oggetti — anticipare la caduta di un oggetto, gestire la dinamica di una superficie deformabile, adattarsi a variazioni nella velocità di movimento — mentre produca prestazioni comparabili o leggermente inferiori nei task che richiedono principalmente categorizzazione semantica degli oggetti statici, dove DINOv2 ha un profilo di competenza già sviluppato.

Il dataset di preaddestramento per seq-JEPA come Perception Module VLA è un problema non banale: le sequenze video egocentriche più rilevanti per il

controllo robotico includono Ego4D (attività umane egocentriche quotidiane) e i video robotici di Open X-Embodiment. La combinazione di questi due corpora — con la differenza sistematica nella prospettiva (ego-centrica umana vs. ego-centrica robotica) e nel tipo di interazione (manipolazione umana fine-grained vs. manipolazione robotica più coarse) — dovrebbe produrre un encoder con un profilo di competenza che unisce la gamma di interazioni fisiche umane con la prospettiva tipica del robot durante il controllo.

La metrica principale del confronto è il tasso di successo sui task di OpenVLA standard — i 29 task distribuiti su più embodiment documentati in Kim et al. (2024) — ma con un'analisi disaggregata per tipologia di task: separare i task di categorizzazione semantica (dove DINOv2 dovrebbe essere più competitivo) dai task di manipolazione fisica dinamica (dove seq-JEPA dovrebbe avere il vantaggio). Questa disaggregazione è cruciale per testare la previsione specifica del programma: non si afferma che seq-JEPA sia uniformemente superiore a DINOv2, ma che sia superiore esattamente nella dimensione dove la sua architettura offre un vantaggio strutturale.

20.8 Sotto-programma VII: la caratterizzazione formale dell'inductive bias di JEPA in termini di strutture causali

Il teorema NFL causale del Capitolo 19 ha mostrato che qualsiasi inductive bias è favorevole per alcune strutture causali e sfavorevole per altre. Un sotto-programma di ricerca fondamentale — più teorico degli altri — è la caratterizzazione formale dell'inductive bias di JEPA in termini di strutture causali, ossia la risposta precisa alla domanda: quali strutture causali favorisce il paradigma JEPA, e quali svantaggia?

La risposta formale richiede la specificazione di una classe di processi generativi causali $\mathcal{C}_{\text{JEPA}} \subseteq \mathcal{F}$ per cui l'obiettivo JEPA è asintoticamente efficiente — ossia per cui il minimo della perdita JEPA coincide con il processo generativo causale vero — e del suo complemento $\mathcal{C}_{\text{JEPA}}^c$ per cui l'obiettivo JEPA non garantisce l'efficienza.

Sulla base dell'analisi del Capitolo 17, si può congetturare che $\mathcal{C}_{\text{JEPA}}$ includa i processi generativi che soddisfano le seguenti condizioni:

C1 — Continuità delle transizioni: la funzione di transizione $f : \mathbb{Z} \times \mathcal{A} \rightarrow \mathbb{Z}$ è continua e differenziabile quasi ovunque nello spazio latente, con costante di Lipschitz finita: $\|f(z, a) - f(z', a)\| \leq L_z \cdot \|z - z'\|$ per ogni z, z', a .

C2 — Struttura markoviana dell'ordine giusto: lo stato latente z_t è markoviano di ordine k per qualche k piccolo — ossia la distribuzione di z_{t+1} condizionata su $(z_t, a_t, z_{t-1}, a_{t-1}, \dots)$ dipende soltanto da $(z_t, z_{t-1}, \dots, z_{t-k+1})$ e da $(a_t, \dots, a_{t-k+1})$.

C3 — Separabilità delle componenti: le variabili causali latenti soddisfano una forma di separabilità — la funzione di transizione è separabile in componenti: $f(z, a) = (f_1(z^{\{1\}}, a_1), f_2(z^{\{2\}}, a_2), \dots, f_d(z^{\{d\}}, a_d))$ per una partizione delle componenti di z e di a .

C4 — Ricchezza della variabilità delle azioni: le azioni coprono sufficientemente lo spazio degli stati da garantire la condizione di sufficienza dell'iVAE — ossia la matrice jacobiana $\partial \eta(a)/\partial a$ ha rango d quasi ovunque nel supporto di $P(a)$.

Queste quattro condizioni definiscono la classe congetturata $\mathcal{C}_{\text{JEPA}}$. La caratterizzazione formale richiede la dimostrazione di due risultati: (1) che per ogni processo in $\mathcal{C}_{\text{JEPA}}$, il minimo dell'obiettivo JEPA è raggiunto da una funzione che recupera i fattori causali (nel senso dell'identificabilità iVAE); e (2) che per ogni processo in $\mathcal{C}_{\text{JEPA}}^c$ (che violi almeno una delle quattro condizioni), esistono funzioni diverse che raggiungono lo stesso minimo dell'obiettivo JEPA pur producendo spazi latenti con strutture causali radicalmente diverse.

Il risultato (1) è parzialmente derivato dalla connessione con il teorema iVAE stabilita nel Capitolo 17, ma richiede una dimostrazione più completa che tenga conto della presenza del stop-gradient e della regolarizzazione SIGReg nell'obiettivo JEPA concreto, non soltanto nell'obiettivo astratto. Il risultato (2) si ottiene attraverso la costruzione di controesempi espliciti per ciascuna violazione delle condizioni C1–C4 — un programma costruttivo che richiede significativa creatività matematica ma non strumenti teorici al di là di quelli già sviluppati nel Causal Representation Learning.

La caratterizzazione formale dell'inductive bias ha conseguenze dirette per l'applicazione pratica: se si sa che il processo generativo di un'applicazione specifica soddisfa le condizioni C1–C4, si può concludere che JEPA è la scelta principata per quel dominio; se si sa che le condizioni sono violate, si può progettare strategie di mitigazione specifiche (per esempio, aggiungere history conditioning per soddisfare C2, o aumentare la diversità delle azioni di training per soddisfare C4).

20.9 Sotto-programma VIII: la pianificazione gerarchica a lungo orizzonte con World Models basati su V-JEPA 2.1

Il Capitolo 11 ha derivato la struttura teorica del reinforcement learning gerarchico con World Models basati su JEPA, identificando come limite attuale la pianificazione a lungo orizzonte in spazi fisici complessi. Un sotto-programma naturale è l'implementazione e valutazione di un sistema gerarchico a due livelli che combina LeWorldModel come World Model di basso livello con un World Model astratto costruito sulle rappresentazioni dense di V-JEPA 2.1.

Il design architetturale proposto per il sistema gerarchico è il seguente. Al livello basso, LeWorldModel opera sulle rappresentazioni latenti di V-JEPA 2.1 con frame skip $k = 5$ (equivalente a 25 step primitivi dell'ambiente), predice il successore latente $\hat{z}_{t+5}$ dato lo stato corrente $\hat{z}_t$ e l'azione a_t , e pianifica con il Cross-Entropy Method su un orizzonte $H_{low} = 5$ step del livello basso. Al livello alto, un World Model astratto $f_{high} : \mathbb{Z}_{agg} \times \Omega \rightarrow \mathbb{Z}_{agg}$ opera sulle rappresentazioni aggregate z_{AGG} di V-JEPA 2.1, predice la transizione tra sotto-obiettivi $\zeta_k = z_{AGG}(o_{tk})$ e $\zeta_{k+1} = z_{AGG}(o_{t(k+1)})$ condizionatamente all'opzione ω_k (un sotto-obiettivo intermedio), e pianifica su un orizzonte $H_{high} = 5$ step del livello alto (equivalente a 25 step del livello basso, ossia 125 step primitivi dell'ambiente).

L'interazione tra i due livelli segue lo schema dell'HRL con options: il planner di alto livello seleziona ogni H_{low} step del livello basso (ogni 5 step del livello alto) un nuovo sotto-obiettivo ζ_{target} nello spazio z_{AGG} ; il planner di basso livello pianifica la sequenza di azioni primitive che minimizza la distanza latente tra $\hat{z}_t$ e ζ_{target} :

$$a^*_{t:t+H_{low}} = \underset{a_{t:t+H_{low}}}{\operatorname{argmin}} \{ \|\hat{z}_{t+H_{low}}(a_{t:t+H_{low}}) - \zeta_{target}\|^2 \}$$

dove $\hat{z}_{t+H_{low}}(a_{t:t+H_{low}})$ è la traiettoria latente prevista da LeWorldModel date le azioni pianificate.

Questo sistema produce una pianificazione a orizzonte effettivo di $5 \times 25 = 125$ step primitivi — un'estensione considerevole rispetto all'orizzonte $H = 5$ step di LeWorldModel baseline — potenzialmente sufficiente per task di manipolazione multi-step come portare un oggetto da una posizione A a una posizione B attraverso una sequenza di sotto-movimenti intermedi, o navigare in un ambiente con ostacoli richiedendo più di un movimento di evitamento. La valutazione su task che richiedono orizzonte lungo — disponibili nel framework OGBench di Ghugare et al. (2024) nelle varianti multi-step — permetterebbe di quantificare il guadagno dell'architettura gerarchica rispetto al baseline a singolo livello.

Il costo computazionale del sistema gerarchico è stimabile analiticamente. Per N_{high} sequenze candidate al livello alto di orizzonte H_{high} e N_{low} sequenze candidate al livello basso di orizzonte H_{low} , il numero totale di valutazioni del World Model è:

$$N_{total} = N_{high} \cdot H_{high} \cdot N_{low} \cdot H_{low}$$

Per $N_{high} = N_{low} = 100$ e $H_{high} = H_{low} = 5$, si hanno $N_{total} = 100 \cdot 5 \cdot 100 \cdot 5 = 250.000$ valutazioni del World Model per ogni step di pianificazione del livello alto. Con un World Model leggero (15M parametri come LeWorldModel), ciascuna valutazione richiede circa 0.1 ms su GPU moderna, per un totale di circa 25 secondi per step di pianificazione di alto livello — non in tempo reale per sistemi robotici a frequenza 6 Hz, ma

potenzialmente trattabile per task di pianificazione offline dove la latenza non è critica.

20.10 Sotto-programma IX: allineamento epistemicamente informato con raccolta di preference data equivariante

Il Capitolo 15 ha proposto la raccolta di *preference data che rispettano l'equivarianza* come strategia per prevenire la degradazione della struttura JEPA durante il fine-tuning con RLHF. Formalizzare e implementare questa strategia è il contenuto del presente sotto-programma.

Il design di raccolta dati proposto opera in tre fasi. La prima fase usa il probing causale del Capitolo 12 per identificare le direzioni dello spazio latente z_AGG che corrispondono a variabili causalmente rilevanti per il task ma causalmente irrilevanti per il valore dell'output (nel senso della preferenza dell'utente). Per un sistema di grasping robotico, queste direzioni sono quelle che corrispondono alla prospettiva della camera — la posizione da cui l'oggetto è osservato — che varia senza influenzare la qualità della presa.

La seconda fase costruisce coppie di training per il modello di ricompensa che *neutralizzano* le variabili equivarianti. Per ogni coppia di output (y_w , y_l) proposta agli annotatori, si presentano le stesse scene osservate da *due* prospettive diverse campionate uniformemente: la prospettiva frontale e una prospettiva ruotata di angolo casuale. Si chiede agli annotatori di valutare la qualità della presa *indipendentemente dalla prospettiva* — una istruzione esplicita che elicit giudizi basati sulla qualità fisica della manipolazione piuttosto che sull'aspetto visivo dell'azione. Formalmente, la loss del modello di ricompensa diventa:

$$\mathcal{L}_{eq_RLHF}(r_\psi) = E_{\{(x,y_w,y_l) \sim D\}} [-\log \sigma(r_\psi(z_AGG(x), y_w) - r_\psi(z_AGG(x), y_l))] + \lambda \cdot E_{\{g \sim G\}} [|r_\psi(z_AGG(x), y_w) - r_\psi(z_AGG(g \cdot x), y_w)|^2]$$

dove il secondo termine è un termine di regolarizzazione che penalizza la dipendenza del modello di ricompensa dalla prospettiva g — imponendo che r_ψ sia invariante alla componente equivariante dello spazio JEPA.

La terza fase valuta il modello di ricompensa così addestrato attraverso il test di *causal robustness* del Capitolo 12: si misura la correlazione tra $r_\psi(z_AGG, y)$ e $r^*(z_AGG, y)$ (il reward gold stimato attraverso valutazione esaustiva su un piccolo dataset con ground truth noto) attraverso diverse prospettive, illuminazioni e posizioni degli oggetti. Se la correlazione è alta e stabile attraverso queste variazioni, il modello di ricompensa ha imparato a misurare la qualità causalmente rilevante dell'azione piuttosto che correlati visivi superficiali.

La connessione con il framework delle etichette complementari di Yin et al. emerge naturalmente: la valutazione equivariante proposta qui è una forma di *partitioned expertise applicata alle variabili trasformazionali* — annotatori che sono esperti in fisica della manipolazione e quindi possono valutare la qualità di una presa indipendentemente dalla prospettiva, anche se non sarebbero in grado di valutare aspetti specialistici come la geometria esatta della forza di presa. La combinazione dei due framework — etichette complementari per la valutazione dell'accuratezza aggregata, data collection equivariante per la qualità del modello di ricompensa — è la forma più completa di scalable oversight epistemicamente informato che emerge dal programma di ricerca complessivo.

20.11 La traiettoria unificante: dall'analisi all'architettura al deployment

I nove sotto-programmi descritti nelle sezioni precedenti sono collegati da una traiettoria logica che va dall'analisi teorica all'architettura sperimentale al deployment in sistemi reali. La traiettoria non è strettamente sequenziale — molti sotto-programmi possono procedere in parallelo — ma ha una struttura gerarchica di dipendenze che è utile rendere esplicita per orientare le priorità.

Il sotto-programma I (CV-JEPA) e il sotto-programma VIII (caratterizzazione dell'inductive bias) sono fondamenti teorici che informano tutti gli altri: il primo fornisce l'architettura di base per l'identificabilità causale, il secondo determina quando quell'architettura è appropriata. I sotto-programmi II (SAE visivi) e IV (verifica formale modulare) sono strumenti di analisi che producono evidenza sulla struttura interna dei sistemi JEPA — evidenza che è necessaria per il design consapevole degli altri componenti. I sotto-programmi III (trasferimento fisico), VI (seq-JEPA come Perception Module VLA) e IX (gerarchico a lungo orizzonte) sono valutazioni empiriche che testano le previsioni della teoria sull'efficacia dei sistemi JEPA in ambienti fisici reali. I sotto-programmi V (oversight ibrido) e IX (allineamento equivariante) sono applicazioni al problema dell'allineamento che si appoggiano sui risultati dei sotto-programmi precedenti.

Il punto di arrivo della traiettoria è un sistema AMI che soddisfa — in misura approssimata e con garanzie quantificate — le tre proprietà dell'ideale regolativo: incarnato nel senso di operare in ambienti fisici reali con osservazioni non pre-processate, verificabile nel senso di permettere un'ispezione strutturata delle proprie rappresentazioni attraverso probing causale e SAE, causalmente coerente nel senso di generalizzare a configurazioni nuove attraverso la struttura d'intervento appresa durante il preaddestramento.

L'aggettivo «approssimata» è fondamentale. Il Capitolo 19 ha documentato con precisione perché l'ideale regolativo non può essere pienamente soddisfatto da nessun sistema finito con supervisione finita in spazi di input-output non banali. Le garanzie quantificate — i bound di deviazione di Yin et al., i result di

probing causale di seq-JEPA, il VoE test di LeWorldModel — forniscono la misura di quanto il sistema approssimi l'ideale, non la prova che lo raggiunga. Questa distinzione non riduce il valore del programma: la precisione con cui si misura l'approssimazione è il contributo che il paradigma JEPA offre al campo dell'intelligenza artificiale incarnata e verificabile. Un sistema il cui comportamento è comprensibile attraverso strumenti teorici espliciti, quantificabile attraverso metriche derivate da principi, e migliorabile attraverso interventi cominciati è già molto più allineato all'ideale regolativo di un sistema le cui proprietà sono empiricamente documentate ma teoricamente opache.

20.12 Nota conclusiva

La monografia si chiude con un'affermazione che è insieme una conclusione e un principio metodologico. Il programma di ricerca delineato in questo capitolo non promette soluzioni complete ai problemi dell'allineamento, della verifica formale o dell'incarnazione computazionale ovvero dell'AGI. Promette qualcosa di più modesto e di più prezioso: un metodo per derivare le domande giuste dalla struttura matematica dei problemi, per misurare con precisione la distanza tra le proprietà desiderate e quelle effettivamente garantibili, e per costruire sistemi il cui comportamento è comprensibile nei termini della teoria che li ha generati.

Il paradigma JEPA — nella sua architettura canonica, nelle sue estensioni variazionali e causali, nella sua incarnazione robotica e nella sua connessione con l'allineamento normativo — è un candidato serio per questo metodo. Non perché risolva tutti i problemi, ma perché la sua struttura matematica è sufficientemente esplicita da permettere la derivazione di previsioni verificabili, la formulazione di esperimenti discriminativi, e il riconoscimento onesto di quando le previsioni non si materializzano. Questa trasparenza epistemica — la capacità di un programma di ricerca di generare le proprie confutazioni — è il criterio che separa la ricerca scientifica dalla speculazione. Il trittico JEPA ha cercato di costruire un programma con questa proprietà. Se ci è riuscito, sarà giudicato dai risultati empirici che i sotto-programmi delineati in questo capitolo produrranno nei prossimi anni.

20.13 Il trittico come programma unitario: architettura d'insieme, divisione del lavoro e tensioni interne

Le tre monografie che compongono il trittico non sono testi paralleli che trattano lo stesso oggetto da angolature diverse: sono livelli sovrapposti di un unico programma di ricerca, ciascuno dei quali presuppone il precedente e ne espande il dominio in una direzione che il precedente non poteva percorrere senza cambiare natura. Riconoscere questa stratificazione non è un gesto

retorico di chiusura: è la condizione per valutare correttamente cosa il programma nel suo insieme ha prodotto e dove si arrestano le sue garanzie.

La prima monografia — quella che qui si chiamerà Rapp-JEPA, seguendo la denominazione adottata dalla stessa Var-T-JEPA — occupa il piano fondazionale nel triplice senso del termine: genealogico, matematico e filosofico. Sul piano genealogico essa ricostruisce la traiettoria intellettuale che conduce da Helmholtz e Piaget alla predictive coding di Rao e Ballard, dai World Models di Schmidhuber alle architetture V-JEPA e LeJEPA, mostrando che ogni passaggio risolve il problema del predecessore al costo di introdurre uno nuovo. Sul piano matematico essa formalizza le proprietà che una rappresentazione latente dovrebbe soddisfare — separabilità semantica, equivarianza alle trasformazioni rilevanti, invarianza a quelle irrilevanti, gaussianità isotrópica come distribuzione ottimale per il probing — e dimostra con rigore (Teorema 1 di LeJEPA) che la distribuzione gaussiana isotropa è l'unico minimizzatore del rischio downstream per qualsiasi task di probing lineare sotto vincoli scalari sulla covarianza. Sul piano filosofico essa non si limita a citare Searle e Chalmers come cautele ornamentali, ma costruisce una stratificazione epistemica articolata in cui diverse nozioni di «comprensione» ricevono risposte diverse, e in cui i teoremi di impossibilità di Putnam, Kripke e Tarski vengono derivati con sufficiente dettaglio da costituire argomenti e non soltanto riferimenti. La funzione di Rapp-JEPA nel trittico è quella di stabilire le coordinate del problema e il vocabolario critico che le monografie successive erediteranno.

La seconda monografia — Var-T-JEPA — scende al piano tecnico-probabilistico e compie un'operazione che Rapp-JEPA aveva solo indicato come necessaria: la reinterpretazione dell'architettura JEPA deterministica come caso limite di un framework variazionale bayesiano. Il contributo di Gögl e Yau (2026) risponde a una domanda di ingegneria inversa probabilistica — quale modello generativo avrebbe come procedura di inferenza naturale esattamente l'architettura JEPA? — con una risposta che è al tempo stesso ovvia in retrospettiva e ricca di conseguenze architettonali. L'ELBO a cinque termini di Var-JEPA non è un'aggiunta tecnica a JEPA: è la rivelazione della struttura generativa che era implicitamente presente fin dall'inizio. Lo stop-gradient e l'EMA dell'encoder target, due scelte che nel JEPA deterministico erano giustificate empiricamente come prevenzione del collasso, emergono nel framework variazionale come conseguenze necessarie della struttura del lower bound: il KL tra il posteriore del target e il prior predittivo è il termine che impedisce il collasso, e il gradiente non scorre attraverso il target encoder perché quest'ultimo parametrizza il prior, non il posteriore dell'inferenza. Var-T-JEPA estende poi questo framework al dominio tabulare — un banco di prova inaspettatamente esigente perché i dati tabulari combinano feature eterogenee (continue, categoriali, ordinali) in distribuzioni altamente asimmetriche — e introduce la valutazione selettiva basata sull'incertezza latente come applicazione pratica della quantificazione bayesiana. LLM-JEPA infine trapianta il principio JEPA nel preaddestramento linguistico

attraverso una perdita composita con predittore a pesi condivisi e maschera di attenzione a blocchi causali, producendo il risultato empirico più sorprendente del trittico: il miglioramento dell'efficienza dei dati di un fattore sedici su task di completamento di espressioni regolari, una violazione delle Scaling Laws di Chinchilla che suggerisce che l'obiettivo standard di predizione del prossimo token lascia inutilizzata una quantità considerevole di struttura informazionale presente nel corpus.

La terza monografia — il presente testo — occupa il piano agentivo-causale: l'integrazione del do-calcolo di Pearl con la struttura generativa di Var-JEPA, l'incarnazione del principio nelle architetture Vision-Language-Action (RT-2, OpenVLA), e la derivazione delle condizioni sotto cui le garanzie di identificabilità causale si trasferiscono dalla teoria alla pratica. Il Capitolo 17 di questa monografia rappresenta il nucleo teorico unificante dell'intero trittico, nell'accezione precisa che ha costruito il primo isomorfismo formale tra la distribuzione interventale del do-calcolo di Pearl e il DAG generativo di Var-JEPA — ovvero tra il livello 2 della gerarchia causale e la struttura dell'ELBO variazionale. Questa connessione non era presente nelle due monografie precedenti: Rapp-JEPA invocava Pearl come orizzonte normativo senza costruire il ponte formale; Var-T-JEPA discuteva la causalità come problema irrisolto nei termini del Capitolo 19.3. È nel presente testo che la connessione viene derivata e i confini del suo dominio di validità ($\mathcal{F}^+$) vengono caratterizzati.

La tensione interna al trittico che merita di essere esplicitata non è di natura tecnica ma epistemica. Rapp-JEPA conclude con un giudizio differenziato sulla plausibilità della via $\text{JEPA} \rightarrow \text{AMI} \rightarrow \text{AGI}$: alta per i componenti individuali, media per l'integrazione, bassa per la risoluzione delle barriere fondamentali, indeterminata per il salto qualitativo verso la comprensione genuina. Var-T-JEPA conclude con un catalogo delle lacune condivise — radicamento simbolico, causalità interventale, incarnazione, memoria, sensibilità agli iperparametri — che nessuno dei tre contributi tecnici (Var-JEPA, LLM-JEPA, Mumuni & Mumuni) riesce a colmare nemmeno parzialmente. Il presente testo, attraverso le derivazioni formali del Capitolo 17 e la sezione 19.7, è riuscito a fondare formalmente la coerenza causale per la classe $\mathcal{F}^+$ degli SCM favorevoli e la verificabilità condizionata alle assunzioni iVAE — ovvero a trasformare due «lacune programmatiche» di Rapp-JEPA e Var-T-JEPA in risultati condizionalmente dimostrati. Ma ha anche reso più precisi e quindi più visibili i bordi del conseguimento: la fondazione vale per $\mathcal{F}^+$, e la caratterizzazione formale di $\mathcal{F}^+$ attraverso il teorema NFL della sezione 19.7 chiarisce che $\mathcal{F}^+$ è un sottoinsieme proprio — probabilmente di misura piccola — di $\mathcal{F}_{\text{p}}$. Il programma è avanzato, ma i suoi confini sono stati più nitidamente marcati proprio attraverso l'avanzamento.

20.14 La mappa topografica delle lacune condivise: cosa nessuno dei tre contributi risolve

Le tre monografie convergono su un insieme di lacune strutturali che, diversamente dalle questioni tecniche aperte (come l'estensione di Var-JEPA alla visione ad alta risoluzione o la costruzione di viste non banali per LLM-JEPA), non sembrano risolvibili attraverso refinement tecnico incrementale. Sono lacune di tipo categoriale, nel senso che la loro persistenza non riflette la mancanza di un algoritmo migliore ma la presenza di vincoli che operano a un livello logico superiore a quello a cui il framework JEPA si rivolge.

La prima e più profonda di queste lacune è il problema del radicamento causale nell'interazione fisica — l'Harnad-Pearl gap, come lo si potrebbe denominare combinando il Symbol Grounding Problem di Harnad con la gerarchia causale di Pearl. Le tre monografie concordano nell'identificare questo come il nodo critico, ma le loro diagnosi si articolano in modo differente. Rapp-JEPA lo pone nei termini dell'embodied cognition: un sistema che apprende il suo Perception Module da dati passivi e il suo World Model da simulatori è nella posizione epistemica di chi ha studiato il nuoto soltanto da libri — può avere rappresentazioni ricche del nuoto senza aver mai sperimentato la resistenza dell'acqua, la coordinazione motoria, il feedback propriocettivo che trasformano la comprensione astratta in comprensione incarnata. Var-T-JEPA lo pone nei termini della struttura matematica delle perdite: l'ELBO di Var-JEPA e la perdita composita di LLM-JEPA ottimizzano distribuzioni condizionali $p(s_y | s_x)$, non distribuzioni interventali $P(s_y | \text{do}(s_x = \tilde{s}_x))$, e la differenza tra le due — espressa formalmente come:

$$p(s_y | s_x) = \int p(s_y | s_x, u) \cdot p(u | s_x) du$$

contro:

$$P(s_y | \text{do}(s_x = \tilde{s}_x)) = \int p(s_y | \tilde{s}_x, u) \cdot p(u) du$$

— è irriducibile finché la variabile ausiliaria z di Var-JEPA, condizionata su s_x attraverso $q_\varphi(z | s_x)$, può assorbire informazioni sia sulle cause di s_x che su quelle di s_y , rendendo impossibile la separazione intervento-osservazione all'interno del framework. Il presente testo risolve formalmente questa lacuna per la classe $\mathcal{F}^+$, ma il costo dell'esclusione è quello di precisare che le strutture causali al di fuori di $\mathcal{F}^+$ — quelle con transizioni discontinue, rumori non-gaussiani, o azioni endogene — rimangono inaccessibili esattamente come lo erano prima della derivazione del Capitolo 17.

La seconda lacuna condivisa è la composizionalità sistematica nel senso di Fodor e Pylyshyn — la capacità di formare nuove combinazioni di concetti in modo produttivo e sistematico, tale per cui la comprensione di "rosso" e

"quadrato" implichi automaticamente la capacità di rappresentare "un quadrato rosso che non si è mai incontrato nell'addestramento". Le rappresentazioni distribuite prodotte da encoder JEPA, per quanto geometricamente strutturate (gaussiane isotrope, a bassa ridondanza grazie a SIGReg), non hanno struttura compositiva esplicita: l'embedding di "un quadrato rosso su un tavolo bianco" non è, in generale, una funzione algebrica semplice degli embedding di "quadrato", "rosso", "tavolo" e "bianco". Rapp-JEPA identifica questa lacuna come la prima delle cinque barriere fondamentali sulla via dell'AGI e osserva che la direzione di ricerca che integra rappresentazioni strutturate — grafi neurali, architetture slot-based, reti relazionali — con il principio JEPA è ancora nella sua infanzia. Var-T-JEPA ne rileva la conseguenza pratica nella sensibilità dell'LLM-JEPA alle distribuzioni out-of-domain: il miglioramento di efficienza dei dati documentato su NL-RX-SYNTH non si trasferisce automaticamente a combinazioni di concetti non presenti nel corpus di addestramento, perché la prossimità geometrica degli embedding nello spazio coseno non garantisce l'estensibilità compositiva. Il presente testo non affronta direttamente questa lacuna, il che la lascia come la più aperta tra quelle del tritico nel dominio agentivo: le architetture VLA come RT-2 e OpenVLA manifestano questa stessa fragilità — il fallimento sistematico nei task multi-step con più di uno o due livelli di indirezione è, nella sua essenza, un fallimento della composizionalità sistematica nel dominio della pianificazione motoria.

La terza lacuna è il dimenticatoio catastrofico nell'apprendimento continuo — il fenomeno documentato da Goodfellow et al. (2013) per cui l'addestramento su un nuovo dataset sovrascrive le strutture sinaptiche che codificano l'apprendimento precedente. Tutte e tre le monografie lo riconoscono come problema aperto. Var-T-JEPA osserva con precisione che la regolarizzazione KL verso prior gaussiani standard in Var-JEPA potrebbe aggravare il dimenticatoio in scenari di apprendimento sequenziale: se un nuovo dataset impone una distribuzione di s_x radicalmente diversa da quella precedente, la pressione verso $\mathcal{N}(0, I)$ forza il modello a adottare la nuova struttura cancellando quella acquisita. Rapp-JEPA colloca il problema nella cornice più ampia dell'agentività a lungo termine: un sistema AMI che interagisce con ambienti successivi deve accumulare competenze senza perdere quelle precedenti, e le soluzioni esistenti — Elastic Weight Consolidation, Progressive Neural Networks, memoria esterna — mitigano il problema in casi specifici ma non lo risolvono nella generalità richiesta. Nel dominio dei sistemi VLA la stessa struttura si manifesta come il catastrophic forgetting che il co-fine-tuning di RT-2 è progettato a mitigare (sezione 6.2): il fatto che la mitigazione funzioni è incoraggiante, ma il fatto che richieda di mantenere attiva $\mathcal{L}_{web}$ per l'intera durata del training — impedendo la specializzazione completa nel dominio robotico — rivela che la soluzione è palliativa, non strutturale.

La quarta lacuna trasversale è la calibrazione dell'incertezza epistemica in modo robusto e certificabile. Var-T-JEPA documenta la promessa di questo fronte — la correlazione di Spearman $\rho = 0.949$ tra l'incertezza latente di Var-JEPA e l'incertezza simulata su MNIST corrotto suggerisce una calibrazione

sorprendentemente buona — ma osserva anche che la valutazione selettiva che l'articolo propone (scartare i campioni più incerti prima del test) è un'applicazione passiva, non un meccanismo attivo di controllo del comportamento. Rapp-JEPA identifica la calibrazione come quarta barriera fondamentale verso l'AGI e nota che i sistemi neurali profondi sono notoriamente sovraconfidenti su distribuzioni simili al training e sottoconfidenti su distribuzioni fuori distribuzione — producendo comportamenti catastrofici in ambienti non previsti. Il presente testo utilizza la calibrazione di Var-T-JEPA come assunzione nella derivazione del bound di subottimalità del pessimismo epistemico (sezione 10.4): l'esistenza del bound dipende da una costante $\kappa > 0$ tale che $\epsilon(z_t, a_t) \leq \kappa \cdot \sigma_{ep}(z_t, a_t)$, e il valore di κ dipende dalla qualità della calibrazione nel regime out-of-distribution. Quanto più κ è grande, tanto meno il bound è utile, perché $\lambda_{ep} = \gamma \kappa_{max} / (1-\gamma)^2$ diventa proporzionalmente più penalizzante. La lacuna della calibrazione non è quindi un problema autonomo: si propaga come fonte di degradazione nei risultati teorici delle sezioni successive, amplificando i propri effetti attraverso la catena delle derivazioni.

20.15 L'AGI come miraggio: struttura formale di una distanza irriducibile

L'impressione che il programma JEPA→AMI→AGI sia più vicino al suo traguardo di quanto non fosse cinque anni fa è corretta in senso tecnico e ingannevole in senso epistemico. È corretta nel senso che i componenti individuali hanno fatto progressi genuini e misurabili: la fondazione normativa della distribuzione ottimale delle rappresentazioni (LeJEPA), l'unificazione del framework contrastivo e predittivo sotto un principio generativo comune (Var-JEPA, con il bridge InfoNCE come caso limite), la dimostrazione che l'obiettivo di predizione del prossimo token lascia inutilizzata struttura informazionale recuperabile con obiettivi più principalizzati (LLM-JEPA), la connessione formale tra distribuzioni interventali e DAG generativo per la classe $\mathcal{F}^+$ (presente testo, Capitolo 17). È ingannevole nel senso che il progresso tecnico sui componenti lascia intatta — e in certi casi rende più visibile — la distanza che separa il conseguimento attuale dalla proprietà che renderebbe il sistema una forma di intelligenza generale nel senso strutturalmente rilevante.

Per dare a questa distanza una struttura formale precisa anziché una formulazione intuitiva, è utile inquadrarla attraverso la gerarchia causale di Pearl, la struttura del teorema NFL della sezione 19.7, il meccanismo del lavaggio epistemico identificato da Rapp-JEPA, e l'argomento modello-teorico di Putnam. Questi quattro strumenti sono complementari e ciascuno illumina una dimensione diversa della stessa irriducibilità.

La gerarchia causale come metro di distanza. Pearl articola la competenza causale in tre livelli: il livello associativo ($P(Y \mid X)$ — la distribuzione condizionale, il regime delle correlazioni statistiche), il livello interventale ($P(Y \mid \text{do}(X = x))$ — la distribuzione sotto intervento, il regime della manipolazione

controllata), e il livello controfattuale ($P(Y_x \mid X = x', Y = y')$) — la distribuzione sotto la domanda «cosa sarebbe accaduto se», il regime del ragionamento ipotetico). I tre livelli sono irriducibili l'uno all'altro: nessuna quantità di osservazioni al livello 1 permette di inferire le distribuzioni del livello 2 senza assunzioni strutturali sul meccanismo causale; nessuna quantità di interventi al livello 2 permette di inferire le distribuzioni del livello 3 senza accesso alla struttura completa dello SCM. Questa irriducibilità non è una limitazione degli algoritmi disponibili: è un teorema che segue dalla struttura logica dei tre livelli.

Il framework JEPA — nella sua versione originale, nelle sue estensioni variazionali (Var-JEPA) e nelle sue applicazioni agentive (VLA) — opera strutturalmente al confine tra il livello 1 e il livello 2. Il risultato del Capitolo 17 di questo testo ha mostrato che, per la classe $\mathcal{F}^*$, il predittore $\text{pred}_\varphi(z_t, a_t)$ apprende direttamente la distribuzione interventale anziché soltanto quella condizionale. Questo è un avanzamento formale genuino: il framework si è spostato dal livello 1 al livello 2 per gli SCM che soddisfano le condizioni di Lipschitz-continuità, gaussianità del rumore ed esogeneità delle azioni. Ma il livello 3 — il ragionamento controfattuale del tipo «cosa sarebbe successo se avessi eseguito l'azione diversa?», che Pearl identifica come la preconditione per la comprensione causale nel senso pieno della parola — rimane al di fuori del dominio formale di JEPA. Non perché JEPA non sia abbastanza potente, ma perché il ragionamento controfattuale richiede l'accesso alla struttura completa dello SCM (le equazioni strutturali f_i , non soltanto la distribuzione marginale delle variabili osservabili), e JEPA — come qualsiasi metodo che apprende da osservazioni, anche interventali — non può costruire questa struttura senza assunzioni identificative aggiuntive che vanno al di là di ciò che le architetture VLA, LLM-JEPA o Var-JEPA specificano. L'identificazione delle equazioni strutturali richiede variabilità contestuale strutturata (condizione iVAE), assunzione di non-linearità controllata, e restrizioni sulle distribuzioni del rumore che Rapp-JEPA e Var-T-JEPA concordano nel riconoscere come non garantite nei domini applicativi reali. L'AGI richiede il livello 3; JEPA ha raggiunto il livello 2 per $\mathcal{F}^*$; la distanza tra i due non è una questione di scala ma di struttura logica.

Il No Free Lunch come vincolo sulla generalità. Il teorema NFL della sezione 19.7 ha una conseguenza che va oltre il piano tecnico della scoperta causale. Nella sua generalità, il NFL afferma che non esiste un inductive bias che migliori le prestazioni mediamente su tutti i problemi — ogni guadagno in un sottoinsieme è bilanciato da una perdita nel complementare. Applicato al programma JEPA→AMI→AGI, questo significa che l'inductive bias specifico di JEPA — continuità delle transizioni, gaussianità dei fattori, esogeneità delle azioni — che produce i suoi vantaggi documentati (identificabilità causale su 3DIEBench, calibrazione dell'incertezza su MNIST corrotto, efficienza dei dati su NL-RX-SYNTH) produce necessariamente svantaggi strutturali su strutture causali che violano queste assunzioni. Un sistema di intelligenza generale deve operare su qualsiasi struttura causale che il mondo fisico presenta, non soltanto

su quelle che appartengono a $\mathcal{F}^+$. La distanza tra $\mathcal{F}^+$ e $\mathcal{F}_{-p}$ non è quindi un parametro da ottimizzare progressivamente avvicinando $\mathcal{F}^+$ a $\mathcal{F}_{-p}$: allargare $\mathcal{F}^+$ significa modificare l'inductive bias, e ogni modifica dell'inductive bias riposiziona la partizione del NFL senza eliminarla. Un AGI che non soffra di nessun NFL è un sistema senza inductive bias, ovvero un sistema che non generalizza — una contraddizione in termini. La questione non è se il programma JEPA possa essere costruito senza NFL, ma se l'inductive bias di JEPA sia abbastanza simile alla struttura causale del mondo reale da rendere $\mathcal{F}_{-p} \setminus \mathcal{F}^+$ trascurabile in pratica. Non esiste al momento nessuna evidenza teorica o empirica che permetterebbe di rispondere affermativamente a questa domanda.

Il lavaggio epistemico come meccanismo di sopravvalutazione sistematica. Rapp-JEPA dedica il Capitolo XI alla descrizione del meccanismo che Nourizadeh (2026) chiama «lavaggio epistemico»: il processo in otto stadi attraverso cui la struttura categoriale umana entra nei sistemi di apprendimento profondo attraverso i dati di training curati, scompare dalla vista durante l'elaborazione formale, e riemerge all'output come capacità emergente attribuita erroneamente al sistema anziché alla struttura ereditata. Il meccanismo si applica con eguale forza a V-JEPA 2 — dove il corpus YT1B viene filtrato attraverso Kinetics, Something-Something v2, COIN e EpicKitchens, tutti dataset che codificano giudizi umani espliciti su quale granularità temporale e categoriale è significativa — e ai sistemi VLA analizzati nel presente testo: il dataset Open X-Embodiment contiene 970 mila episodi raccolti da esseri umani che eseguono task selezionati da esseri umani con robot controllati da esseri umani in ambienti progettati da esseri umani. La straordinaria capacità di generalizzazione di RT-2-PaLI-X-55B nei test out-of-distribution non è, in questo senso, la prova che il sistema ha appreso la struttura causale del mondo dalla propria esperienza: è la prova che la struttura causale del mondo era già contenuta, in forma codificata, nella straordinaria quantità di struttura umana incorporata nei dati di preaddestramento e di co-fine-tuning. Questo non diminuisce il valore ingegneristico del sistema — la capacità di estrarre e generalizzare la struttura umana incorporata nei dati è essa stessa una competenza di rilievo — ma rende infondata la traslazione diretta da «il sistema generalizza bene» a «il sistema comprende il mondo». La prima affermazione è verificata; la seconda non è verificata e non è verificabile attraverso i benchmark disponibili, perché i benchmark stessi misurano la generalizzazione rispetto alla struttura umana che è già stata incorporata nei dati.

L'argomento modello-teorico di Putnam come limite semantico irriducibile. L'argomento modello-teorico che Hilary Putnam ha formulato in *Reason, Truth and History* (1981) applicando il teorema di Löwenheim-Skolem alla semantica computazionale stabilisce che qualsiasi teoria formalmente consistente — e in particolare qualsiasi sistema di apprendimento profondo che ottimizza una funzione obiettivo — ammette infiniti modelli distinti rispetto a qualsiasi interpretazione semantica, e non esiste alcun vincolo interno al sistema che permetta di selezionare il modello «corretto» tra questi infiniti. La

conseguenza per il presente contesto è precisa: anche un sistema JEPA che soddisfi tutte le condizioni di $\mathcal{F}^*$, identifichi correttamente le equazioni strutturali degli SCM nel proprio dominio applicativo, e predica le distribuzioni interventali con zero errore, non avrebbe ancora ragione di ritenere che i propri fattori latenti $z_1, z_2, \dots, z_d$ corrispondano ai fattori causali del mondo fisico piuttosto che a una qualsiasi altra struttura algebrica che soddisfi le stesse equazioni. Le coordinate dello spazio latente non hanno referenti fisici determinati dall'interno del sistema: li ricevono, se li ricevono, attraverso la supervisione esterna (le etichette di probing, i task downstream, le valutazioni umane). La biezione garantita dall'iVAE tra z e i fattori causali veri S è una biezione *a meno di trasformazione elementwise* — il che significa che c'è ancora una famiglia infinita di bijections compatibili con le osservazioni, e senza conoscenza esterna del dominio non è possibile distinguerle. L'AGI come sistema autonomo che comprende il mondo fisico dalla propria esperienza senza supervisione esterna non può essere fondato su un framework che richiede supervisione esterna per ancorare semanticamente le proprie rappresentazioni.

Cosa resta aperto e cosa non è mai stato chiuso. La distinzione tra «lacune tecniche aperte» e «limiti strutturali irriducibili» attraversa il trittico senza essere mai resa completamente esplicita in un unico luogo. Una formulazione precisa sarebbe la seguente. Sono lacune tecniche aperte — problemi che non sono stati risolti ma per cui esistono strategie di ricerca identificabili con possibilità non nulla di successo — l'estensione di Var-JEPA alla visione ad alta risoluzione, la costruzione di viste non banali per LLM-JEPA nel dominio del linguaggio arbitrario, la riduzione dell'overhead computazionale del framework variazionale su sequenze video lunghe, e il raffinamento dell'inductive bias JEPA verso strutture causali con discontinuità moderate (soglie morbide, transizioni quasi-continue). Sono limiti strutturali irriducibili — vincoli che operano a un livello logico che nessun progresso tecnico può rimuovere senza cambiare il problema — l'irriducibilità dei tre livelli della gerarchia causale di Pearl (e in particolare l'impossibilità di inferire il livello 3 dal livello 2 senza struttura completa dell'SCM), il teorema NFL (e l'impossibilità di eliminare l'inductive bias senza eliminare la generalizzazione), la non-unicità del modello semantico di Putnam (e l'impossibilità di ancorare il riferimento dall'interno del formalismo), e il Problema Difficile della Coscienza di Chalmers (e l'impossibilità di determinare dall'esterno se un sistema funzionalmente capace possieda esperienza soggettiva nel senso di Nagel e Chalmers).

L'AGI, nel senso in cui questa espressione è usata nel dibattito corrente — un sistema capace di ragionamento causale a livello 3, di comprensione compositiva sistematica, di radicamento simbolico nel mondo fisico, di apprendimento continuo senza dimenticatoio catastrofico, e di agentività con obiettivi autonomi — non è un miraggio nel senso di essere impossibile o privo di interesse, ma è un miraggio nel senso che la distanza da esso non si riduce proporzionalmente all'avanzamento tecnico sui componenti individuali. Per ogni passo che avvicina il programma JEPA ai livelli 2 della gerarchia causale su

un sottoinsieme di strutture, il passo successivo verso il livello 3 sull'intera $\mathcal{F}_p$ non diventa più corto: rimane separato da una discontinuità logica che non si attraversa per continuazione dell'avanzamento corrente, ma richiede la costruzione di qualcosa strutturalmente diverso — che probabilmente assomiglierà al framework JEPA, ma che non può essere costruito incrementalmente da esso.

Questo giudizio non è pessimismo tecnologico. È la forma più onesta di rispetto per il programma di ricerca che il trittico ha delineato: un programma che ha prodotto risultati formalmente dimostrati di rilievo, che ha costruito fondamenta teoriche più solide di qualsiasi predecessore nel dominio dell'apprendimento delle rappresentazioni causali, e che ha identificato con precisione — attraverso la sezione 19.7, il Capitolo 17, e le analisi comparative di questa sezione conclusiva — dove finiscono le garanzie e dove inizia l'aspirazione. La via JEPA→AMI→AGI rimane teoricamente fondata, empiricamente promettente, e filosoficamente onesta sui propri limiti. L'AGI che essa promette è a distanza indeterminata — non perché il programma sia sbagliato, ma perché la distanza verso i limiti strutturali non è espressa in unità di progresso tecnico, e nessun sistema di coordinate attualmente disponibile permette di misurarla.

Appendice A — Derivazione completa dell'ELBO causale di CV-JEPA

A.1 Il modello generativo di partenza

Il framework CV-JEPA riposa su un modello generativo a variabili latenti strutturato secondo tre strati di astrazione. Al livello più profondo risiedono i fattori causali condivisi $s \in \mathbb{R}^m$, distribuiti secondo un prior fattorizzato $p(s) = \prod_i p(s_i)$, il cui significato è quello di catturare i meccanismi stabili che generano simultaneamente la vista contestuale e la vista target. Al livello intermedio risiedono i fattori vista-specifici $c \in \mathbb{R}^d$ e $t \in \mathbb{R}^d$, che codificano le variazioni idiosincratiche di ciascuna vista — variazioni dovute a prospettiva, illuminazione, occlusione parziale, rumore sensoriale — e che il modello generativo genera condizionatamente ai fattori causali condivisi attraverso le equazioni strutturali lineari $c = A_c \cdot s + \epsilon_c$ e $t = A_t \cdot s + \epsilon_t$, dove $A_c \in \mathbb{R}^{d \times m}$ e $A_t \in \mathbb{R}^{d \times m}$ sono le matrici di connessione e ϵ_c, ϵ_t sono rumori isotropi gaussiani indipendenti. Al livello più superficiale si trovano le osservazioni $x_c \in \mathcal{X}$ e $x_t \in \mathcal{X}$, generate a partire dai fattori vista-specifici attraverso decoder parametrizzati: $x_c \sim p_\theta(\cdot | c)$ e $x_t \sim p_\theta(\cdot | t)$.

La struttura fattorizzata del modello generativo congiunto è dunque:

$$p_{\theta}(x_c, x_t, c, t, s) = p_{\theta}(x_c | c) \cdot p_{\theta}(x_t | t) \cdot p_A(c | s) \cdot p_A(t | s) \cdot p(s)$$

dove il pedice A in p_A sottolinea che la dipendenza di c e t da s è mediata dalle matrici di connessione strutturale A_c e A_t , le quali devono soddisfare un vincolo di aciclicità che sarà introdotto nel termine T_7 dell'ELBO. Le distribuzioni condizionali sono:

$$\begin{aligned} p_A(c | s) &= \mathcal{N}(c; A_c \cdot s, \sigma_c^2 I_d) \\ p_A(t | s) &= \mathcal{N}(t; A_t \cdot s, \sigma_t^2 I_d) \\ p(s) &= \prod_{i=1}^m p(s_i) \quad \text{con } p(s_i) \text{ tipicamente } \mathcal{N}(0, 1) \text{ o } \text{Laplace}(0, 1) \end{aligned}$$

La quantità che il modello è chiamato a massimizzare è la log-verosimiglianza marginale delle osservazioni:

$$\log p_{\theta}(x_c, x_t) = \log \iiint p_{\theta}(x_c, x_t, c, t, s) ds dc dt$$

Questa integrale è in generale intrattabile per decoder p_{θ} non-lineari, il che motiva l'introduzione di un lower bound variazionale.

A.2 Il posteriore variazionale fattorizzato

Si introduce la famiglia variazionale $q_{\varphi}(s, c, t | x_c, x_t)$ che approssima il posteriore esatto $p_{\theta}(s, c, t | x_c, x_t)$. Per rendere computazionalmente trattabile il calcolo dell'ELBO si assume la fattorizzazione:

$$q_{\varphi}(s, c, t | x_c, x_t) = q_{\varphi^s}(s | z_c, z_t) \cdot q_{\varphi^c}(c | x_c) \cdot q_{\varphi^t}(t | x_t)$$

dove $z_c = \text{enc}_c(x_c)$ e $z_t = \text{enc}_t(x_t)$ sono le uscite degli encoder deterministici del contesto e del target, rispettivamente. La scelta di condizionare q_{φ^s} sulle rappresentazioni latenti codificate z_c e z_t anziché direttamente sulle osservazioni grezze x_c e x_t è motivata sia computazionalmente — evita di dover passare immagini di alta dimensionalità al modulo di stima dei fattori causali — sia architetturealmente: la struttura di CV-JEPA prevede che i fattori causali emergano dall'interazione tra i due spazi latenti, non dalla sola osservazione di ciascuna vista.

La distribuzione variazionale sui fattori vista-specifici è gaussiana diagonale:

$$\begin{aligned} q_{\varphi^c}(c | x_c) &= \mathcal{N}(c; \mu_c(x_c), \text{diag}(\sigma_c^2(x_c))) \\ q_{\varphi^t}(t | x_t) &= \mathcal{N}(t; \mu_t(x_t), \text{diag}(\sigma_t^2(x_t))) \end{aligned}$$

dove $\mu_c, \sigma_c, \mu_t, \sigma_t$ sono reti neurali con parametri φ . La distribuzione variazionale sui fattori causali condivisi è anch'essa gaussiana diagonale:

$$q_\varphi(s \mid z_c, z_t) = \mathcal{N}(s; \mu_s(z_c, z_t), \text{diag}(\sigma_s^2(z_c, z_t)))$$

La fattorizzazione del posteriore variazionale introduce indipendenza condizionale tra i fattori causali s e i fattori vista-specifici c, t — un'assunzione che riflette la struttura del modello generativo in cui s è la causa comune di c e t , e c e t sono condizionalmente indipendenti dato s .

A.3 Il lower bound variazionale: applicazione della disuguaglianza di Jensen

La derivazione del lower bound variazionale (Evidence Lower BOund, ELBO) procede attraverso l'applicazione della disuguaglianza di Jensen alla log-verosimiglianza marginale. Poiché il logaritmo è una funzione strettamente concava, per ogni distribuzione q_φ si ha:

$$\begin{aligned} \log p_\theta(x_c, x_t) &= \log \iiint p_\theta(x_c, x_t, c, t, s) ds dc dt \\ &= \log \iiint [p_\theta(x_c, x_t, c, t, s) / q_\varphi(s, c, t \mid x_c, x_t)] \cdot q_\varphi(s, c, t \mid x_c, x_t) ds \\ &\quad dc dt \\ &= \log \mathbb{E}_{q_\varphi} \{ [p_\theta(x_c, x_t, c, t, s) / q_\varphi(s, c, t \mid x_c, x_t)] \} \\ &\geq \mathbb{E}_{q_\varphi} \{ [\log p_\theta(x_c, x_t, c, t, s) - \log q_\varphi(s, c, t \mid x_c, x_t)] \} \quad (\text{Jensen}) \end{aligned}$$

Il lower bound è dunque:

$$\mathcal{L}_{\text{CV-JEPA}}(\varphi, \theta, \psi) = \mathbb{E}_{q_\varphi} \{ [\log p_\theta(x_c, x_t, c, t, s) - \log q_\varphi(s, c, t \mid x_c, x_t)] \}$$

con uguaglianza se e solo se $q_\varphi(s, c, t \mid x_c, x_t) = p_\theta(s, c, t \mid x_c, x_t)$ quasi ovunque, ossia quando la distribuzione variazionale coincide esattamente con il posteriore esatto.

Il gap tra la log-verosimiglianza e l'ELBO è esattamente la divergenza di Kullback-Leibler tra il posteriore variazionale e il posteriore esatto:

$$\log p_\theta(x_c, x_t) - \mathcal{L}_{\text{CV-JEPA}} = D_{\text{KL}}(q_\varphi(s, c, t \mid x_c, x_t) \parallel p_\theta(s, c, t \mid x_c, x_t)) \geq 0$$

che è non-negativa per la non-negatività della KL, confermando che $\mathcal{L}_{CV-JEPA}$ è effettivamente un lower bound della log-verosimiglianza marginale.

A.4 Espansione del logaritmo della congiunta

Si espande ora il logaritmo della distribuzione congiunta del modello sfruttando la fattorizzazione $p_{\theta}(x_c, x_t, c, t, s) = p_{\theta}(x_c | c) \cdot p_{\theta}(x_t | t) \cdot p_A(c | s) \cdot p_A(t | s) \cdot p(s)$:

$$\begin{aligned} \log p_{\theta}(x_c, x_t, c, t, s) \\ = \log p_{\theta}(x_c | c) + \log p_{\theta}(x_t | t) + \log p_A(c | s) + \log p_A(t | s) + \log p(s) \end{aligned}$$

Analogamente si espande il logaritmo della distribuzione variazionale usando la fattorizzazione $q_{\varphi} = q_{\varphi^s}(s | z_c, z_t) \cdot q_{\varphi^c}(c | x_c) \cdot q_{\varphi^t}(t | x_t)$:

$$\begin{aligned} \log q_{\varphi}(s, c, t | x_c, x_t) \\ = \log q_{\varphi^s}(s | z_c, z_t) + \log q_{\varphi^c}(c | x_c) + \log q_{\varphi^t}(t | x_t) \end{aligned}$$

L'ELBO diventa:

$$\begin{aligned} \mathcal{L}_{CV-JEPA} = \mathbb{E}_{q_{\varphi}} [\log p_{\theta}(x_c | c)] \\ + \mathbb{E}_{q_{\varphi}} [\log p_{\theta}(x_t | t)] \\ + \mathbb{E}_{q_{\varphi}} [\log p_A(c | s) - \log q_{\varphi^c}(c | x_c)] \\ + \mathbb{E}_{q_{\varphi}} [\log p_A(t | s) - \log q_{\varphi^t}(t | x_t)] \\ + \mathbb{E}_{q_{\varphi}} [\log p(s) - \log q_{\varphi^s}(s | z_c, z_t)] \end{aligned}$$

A.5 Riconoscimento delle divergenze KL

I termini dal terzo al quinto dell'espansione precedente si riconoscono come divergenze di Kullback-Leibler con segno invertito. Ricordando che per distribuzioni P e Q vale:

$$D_{KL}(Q \parallel P) = \mathbb{E}_Q [\log Q - \log P]$$

si ha:

$$\begin{aligned} \mathbb{E}_{q_{\varphi}} [\log p_A(c | s) - \log q_{\varphi^c}(c | x_c)] \\ = \mathbb{E}_{q_{\varphi^s}} [\mathbb{E}_{q_{\varphi^c}} [\log p_A(c | s) - \log q_{\varphi^c}(c | x_c)]] \end{aligned}$$

$$= -\mathbb{E}_{\{q_{\psi^s}\}} [D_{\text{KL}}(q_{\psi^c}(c \mid x_c) \parallel p_A(c \mid s))]$$

dove si è usata la fattorizzazione del posteriore variazionale per separare la media su s dalla media su c . Analogamente:

$$\begin{aligned} & \mathbb{E}_{\{q_{\psi}\}} [\log p_A(t \mid s) - \log q_{\psi^t}(t \mid x_t)] \\ &= -\mathbb{E}_{\{q_{\psi^s}\}} [D_{\text{KL}}(q_{\psi^t}(t \mid x_t) \parallel p_A(t \mid s))] \end{aligned}$$

e per il termine sui fattori causali:

$$\begin{aligned} & \mathbb{E}_{\{q_{\psi}\}} [\log p(s) - \log q_{\psi^s}(s \mid z_c, z_t)] \\ &= -D_{\text{KL}}(q_{\psi^s}(s \mid z_c, z_t) \parallel p(s)) \end{aligned}$$

dove la media è soltanto su q_{ψ^s} poiché né $p(s)$ né q_{ψ^s} dipendono da c o t .

A.6 Introduzione del termine predittivo JEPA

L'ELBO derivato nei paragrafi precedenti massimizza la verosimiglianza e regolarizza le rappresentazioni latenti, ma non impone alcun vincolo sulla capacità del predittore s_{ψ} di prevedere la rappresentazione del target a partire da quella del contesto — il nucleo computazionale dell'architettura JEPA. Questo vincolo si introduce aggiungendo un termine che penalizza la distanza tra la distribuzione variazionale sul target $q_{\psi^t}(t \mid x_t)$ e la distribuzione prevista dal predittore $q_{\psi}(t \mid z_c)$:

$$T_6 = -\mathbb{E}_{\{q_{\psi}\}} [D_{\text{KL}}(q_{\psi^t}(t \mid x_t) \parallel q_{\psi}(t \mid z_c))]$$

La giustificazione probabilistica di T_6 è la seguente: se il predittore $q_{\psi}(t \mid z_c)$ è un buon modello della distribuzione interventale del target dato il contesto, allora la KL tra la distribuzione variazionale reale $q_{\psi^t}(t \mid x_t)$ e quella prevista $q_{\psi}(t \mid z_c)$ dovrebbe essere piccola. Minimizzare la KL equivale a massimizzare la log-verosimiglianza dei fattori target sotto il modello predittivo, il che forza il predittore ad apprendere la distribuzione delle transizioni causali, non soltanto le correlazioni osservative.

La distinzione tra $q_{\psi^t}(t \mid x_t)$ — il posteriore variazionale che usa l'osservazione target effettiva — e $q_{\psi}(t \mid z_c)$ — la predizione basata sul solo contesto — è essenziale: il meccanismo di stop-gradient, già presente nell'architettura originale JEPA di Assran et al. e formalizzato nell'ELBO di Var-JEPA da Gögl e Yau, blocca il flusso del gradiente attraverso il termine T_6 verso l'encoder del target, prevenendo il collasso della soluzione banale in cui il predittore apprende a copiare piuttosto che a prevedere.

A.7 Il vincolo di aciclicità NOTEARS

Il settimo e ultimo termine dell'ELBO causale non emerge direttamente dalla disuguaglianza di Jensen ma viene aggiunto come vincolo lagrangiano al problema di ottimizzazione per garantire che le matrici di connessione A_c e A_t definiscano strutture acicliche — condizione necessaria affinché il modello generativo CV-JEPA abbia un'interpretazione causale consistente.

La funzione $h(A)$ imposta da Zheng et al. nel framework NOTEARS soddisfa:

$h(A) = \text{tr}(e^{\{A \circ A\}}) - d = 0 \iff$ il grafo con matrice di adiacenza $(A \circ A)$ è aciclico

dove $A \circ A$ denota il prodotto di Hadamard elemento per elemento, e $\text{tr}(e^B) = \sum_{k=0}^{\infty} \text{tr}(B^k)/k!$ è la traccia dell'esponenziale matriciale. La equivalenza si dimostra osservando che $\text{tr}((A \circ A)^k)$ conta i cammini di lunghezza k nel grafo con pesi $|A_{ij}|^2$, e dunque $\text{tr}(e^{\{A \circ A\}}) = d + \sum_{k=1}^{\infty} \text{tr}((A \circ A)^k)/k!$ è strettamente maggiore di d se e solo se il grafo ammette cicli. Il costo computazionale del calcolo di $h(A)$ è $O(d^3)$ per la diagonalizzazione della matrice $d \times d$, dove d è la dimensionalità dello spazio latente causale.

Il termine di penalità nell'ELBO è:

$T_7 = -\gamma \cdot h(A_c, A_t)$ dove $h(A_c, A_t) = h(A_c) + h(A_t)$

Il parametro $\gamma \geq 0$ è moltiplicatore di Lagrange che viene incrementato progressivamente durante l'addestramento secondo il metodo del Lagrangiano aumentato, garantendo la convergenza verso soluzioni con $h(A_c, A_t) = 0$ a patto che l'ottimizzazione raggiunga un punto stazionario del problema penalizzato.

A.8 L'ELBO causale completo a sette termini

Raccogliendo i contributi dei paragrafi precedenti, l'ELBO causale di CV-JEPA nella sua forma completa è:

$\mathcal{L}_{\text{CV-JEPA}}(\varphi, \theta, \psi) =$

$T_1 + \mathbb{E}_{\{q_\varphi\}} [\log p_\theta(x_c | c)]$ ricostruzione contesto

$T_2 + \mathbb{E}_{\{q_\varphi\}} [\log p_\theta(x_t | t)]$ ricostruzione target

$$T_3 - \mathbb{E}_{\{q_{\varphi^s}\}} [D_{\text{KL}}(q_{\varphi^c}(c \mid x_c) \parallel p_A(c \mid s))] \quad \text{regolarizzazione } c \rightarrow s$$

$$T_4 - \mathbb{E}_{\{q_{\varphi^s}\}} [D_{\text{KL}}(q_{\varphi^t}(t \mid x_t) \parallel p_A(t \mid s))] \quad \text{regolarizzazione } t \rightarrow s$$

$$T_5 - D_{\text{KL}}(q_{\varphi^s}(s \mid z_c, z_t) \parallel p(s)) \quad \text{regolarizzazione } s \rightarrow \text{prior}$$

$$T_6 - \mathbb{E}_{\{q_{\varphi}\}} [D_{\text{KL}}(q_{\varphi^t}(t \mid x_t) \parallel q_{\psi}(t \mid z_c))] \quad \text{errore predittivo JEPA}$$

$$T_7 - \gamma \cdot [h(A_c) + h(A_t)] \quad \text{vincolo di aciclicità}$$

Questa è la forma esatta dell'ELBO causale di CV-JEPA. Ogni termine discende direttamente dal modello generativo: T_1 e T_2 dalla verosimiglianza dei decoder, T_3 e T_4 dalla condizionale strutturale $p_A(c \mid s)$ e $p_A(t \mid s)$, T_5 dalla regolarizzazione del prior sui fattori causali, T_6 dall'obiettivo di previsione JEPA, T_7 dal vincolo causale sul DAG latente.

In forma compatta, con la notazione dell'ELBO surgery di Mathieu et al., si può scrivere:

$$\mathcal{L}_{\text{CV-JEPA}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{regol}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{causal}}$$

dove:

$$\begin{aligned} \mathcal{L}_{\text{recon}} &= \mathbb{E}[\log p_{\theta}(x_c \mid c)] + \mathbb{E}[\log p_{\theta}(x_t \mid t)] \\ \mathcal{L}_{\text{regol}} &= -\mathbb{E}_{\{q_{\varphi^s}\}} [D_{\text{KL}}(q_{\varphi^c} \parallel p_A(c \mid s))] - \mathbb{E}_{\{q_{\varphi^s}\}} [D_{\text{KL}}(q_{\varphi^t} \parallel p_A(t \mid s))] - D_{\text{KL}}(q_{\varphi^s} \parallel p(s)) \\ \mathcal{L}_{\text{pred}} &= -\mathbb{E}_{\{q_{\varphi}\}} [D_{\text{KL}}(q_{\varphi^t}(t \mid x_t) \parallel q_{\psi}(t \mid z_c))] \\ \mathcal{L}_{\text{causal}} &= -\gamma \cdot [h(A_c) + h(A_t)] \end{aligned}$$

A.9 Relazione con l'ELBO a cinque termini di Var-JEPA

L'ELBO di Var-JEPA derivato da Gögl e Yau si articola in cinque termini che è utile confrontare puntualmente con i sette termini di CV-JEPA per verificare che quest'ultimo ne sia un'estensione strutturalmente coerente e non una proposta assemblata in modo arbitrario.

L'ELBO di Var-JEPA per una coppia (x_c, x_t) è:

$$\mathcal{L}_{\text{Var-JEPA}} = \mathbb{E}_{\{q_{\varphi}\}} [\log p_{\theta}(x_c \mid z_c)] \quad (\text{R1})$$

$$\begin{aligned}
& - D_{\text{KL}}(q_{\psi}(z_c | x_c) \parallel p(z_c)) & (\text{KL1}) \\
& - D_{\text{KL}}(q_{\psi}(z_t | x_t) \parallel p(z_t)) & (\text{KL2}) \\
& - \mathbb{E}_{\{q_{\psi}\}} [D_{\text{KL}}(q_{\psi}(z_t | x_t) \parallel q_{\psi}(z_t | z_c))] & (\text{PRED}) \\
& - \beta \cdot D_{\text{KL}}(q_{\psi}(z_t | x_t) \parallel p(z_t | z_c)) & (\text{PRIOR})
\end{aligned}$$

La corrispondenza con i termini di CV-JEPA è la seguente. Il termine R1 di Var-JEPA corrisponde a $T_1 + T_2$ di CV-JEPA: Var-JEPA ha un unico termine di ricostruzione (tipicamente quello del contesto, con il target ricostruito implicitamente attraverso il predittore), mentre CV-JEPA esplicita la ricostruzione di entrambe le viste. I termini KL1 e KL2 di Var-JEPA corrispondono congiuntamente a $T_3 + T_4 + T_5$ di CV-JEPA: in Var-JEPA i fattori causali condivisi s non sono esplicitati e le regolarizzazioni KL agiscono direttamente sulle rappresentazioni latenti z_c e z_t , mentre in CV-JEPA la gerarchia a tre livelli (s , c e t separati) introduce una KL aggiuntiva T_5 sui fattori causali. Il termine PRED di Var-JEPA corrisponde esattamente a T_6 di CV-JEPA. Il termine PRIOR di Var-JEPA non ha corrispettivo diretto in CV-JEPA: il suo ruolo è svolto dai termini T_3 e T_4 che regolarizzano c e t verso i loro prior condizionali $p_A(c|s)$ e $p_A(t|s)$. Il termine T_7 di CV-JEPA è l'unico genuinamente nuovo: non ha omologo in Var-JEPA e riflette il vincolo causale che distingue CV-JEPA da un semplice VAE gerarchico.

Il passaggio da cinque a sette termini riflette dunque l'arricchimento del modello generativo: l'introduzione esplicita del livello causale s e del vincolo di aciclicità. Ogni termine aggiuntivo è giustificato dalla struttura del modello, non aggiunto euristicamente.

A.10 Calcolo esplicito delle KL gaussiane

Poiché tutte le distribuzioni variazionali e i prior condizionali sono gaussiani, le divergenze KL in T_3 , T_4 e T_5 ammettono espressioni analitiche che si calcolano senza richiedere campionamento Monte Carlo. Per due gaussiane multivariate di dimensione d :

$$\begin{aligned}
D_{\text{KL}}(\mathcal{N}(\mu_1, \Sigma_1) \parallel \mathcal{N}(\mu_2, \Sigma_2)) = \\
\frac{1}{2} [\text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - d + \log(\det(\Sigma_2)/\det(\Sigma_1))]
\end{aligned}$$

Per il termine T_5 , con $q_{\psi}(s | z_c, z_t) = \mathcal{N}(\mu_s, \text{diag}(\sigma_s^2))$ e $p(s) = \mathcal{N}(0, I_m)$:

$$D_{\text{KL}}(q_{\psi} \parallel p(s)) = \frac{1}{2} \sum_{i=1}^m (\mu_{s,i}^2 + \sigma_{s,i}^2 - 1 - \log \sigma_{s,i}^2)$$

questa è la forma standard della KL gaussiana diagonale verso il prior standard, utilizzata in ogni VAE.

Per il termine T_3 , con $q_{\varphi^c}(c \mid x_c) = \mathcal{N}(\mu_c, \text{diag}(\sigma_c^2))$ e $p_A(c \mid s) = \mathcal{N}(A_c \cdot s, \sigma_c^2 I_d)$:

$$D_{\text{KL}}(q_{\varphi^c}(c \mid x_c) \parallel p_A(c \mid s)) = \frac{1}{2} [d \cdot \sigma_c^2(x_c) / \sigma_c^2 + \|\mu_c(x_c) - A_c \cdot s\|^2 / \sigma_c^2 - d + d \log(\sigma_c^2 / \sigma_c^2(x_c))]$$

dove il primo addendo è il termine di varianza, il secondo è il termine di media che penalizza la distanza tra la rappresentazione encodificata $\mu_c(x_c)$ e la predizione lineare $A_c \cdot s$ dei fattori causali, e il terzo è il termine logaritmico del rapporto delle varianze. Questo termine realizza il collegamento strutturale tra i fattori vista-specifici c e i fattori causali condivisi s : se $A_c \cdot s$ è lontano da $\mu_c(x_c)$, il termine T_3 è grande e penalizza la soluzione. Il termine T_4 si calcola analogamente con le quantità corrispondenti al target.

A.11 Stima Monte Carlo e il reparametrization trick

I termini T_1 , T_2 e T_6 dell'ELBO richiedono la valutazione di aspettative rispetto alle distribuzioni variazionali q_{φ^c} , q_{φ^t} , q_{φ^s} che non ammettono forma analitica per decoder p_{θ} non-lineari. Si ricorre alla stima Monte Carlo combinata con il reparametrization trick di Kingma e Welling per rendere il gradiente dell'ELBO rispetto ai parametri φ calcolabile con retropropagazione standard.

Il trick di reparametrizzazione sostituisce il campionamento $z \sim \mathcal{N}(\mu, \sigma^2)$ — operazione non differenziabile rispetto a μ e σ — con la trasformazione equivalente $z = \mu + \sigma \odot \epsilon$ dove $\epsilon \sim \mathcal{N}(0, I)$ è un rumore ausiliario fisso e $\odot$ denota il prodotto di Hadamard. In questo modo:

$$\mathbb{E}_{\{q_{\varphi^c}(c \mid x_c)\}} [f(c)] = \mathbb{E}_{\{\epsilon \sim \mathcal{N}(0, I)\}} [f(\mu_c(x_c) + \sigma_c(x_c) \odot \epsilon)]$$

e il gradiente rispetto a φ può essere trasportato all'interno dell'aspettazione, producendo uno stimatore con varianza bassa.

Con L campioni $\epsilon^{(1)}, \dots, \epsilon^{(L)} \sim \mathcal{N}(0, I)$ lo stimatore a campione singolo ($L=1$) sufficiente in pratica è:

$$T_1 \approx \log p_{\theta}(x_c \mid \mu_c(x_c) + \sigma_c(x_c) \odot \epsilon) \quad \text{con } \epsilon \sim \mathcal{N}(0, I)$$

$$T_6 \approx -D_{\text{KL}}(q_{\varphi^t}(t \mid x_t) \parallel q_{\varphi}(t \mid z_c))$$

dove il termine T_6 ammette calcolo analitico se anche q_ψ è gaussiana.

A.12 Il metodo del Lagrangiano aumentato per il vincolo di aciclicità

Il vincolo $h(A_c) + h(A_t) = 0$ in T_7 è un vincolo di uguaglianza non convesso che non può essere gestito con semplice penalizzazione quadratica, la quale non garantisce la convergenza esatta al vincolo ma soltanto una soluzione approssimata con h proporzionale all'inverso del peso di penalità.

Il metodo del Lagrangiano aumentato (Lagrangian augmented, o metodo dei moltiplicatori) di Hestenes e Powell gestisce questo vincolo in modo esatto attraverso la sequenza di sotto-problemi non vincolati:

$$\min_{\{A_c, A_t, \varphi, \theta, \psi\}} \mathcal{L}_{\text{aumentato}}(\varphi, \theta, \psi, A_c, A_t; \lambda, \mu) = -\mathcal{L}_{\text{CV-JEPA}}(\varphi, \theta, \psi) + \lambda \cdot [h(A_c) + h(A_t)] + (\mu/2) \cdot [h(A_c) + h(A_t)]^2$$

dove $\lambda \in \mathbb{R}$ è il moltiplicatore di Lagrange corrente e $\mu > 0$ è il coefficiente di penalità che viene incrementato progressivamente. L'aggiornamento del moltiplicatore segue:

$$\lambda \leftarrow \lambda + \mu \cdot [h(A_c) + h(A_t)]$$

a ogni passo esterno dell'ottimizzazione. La convergenza al vincolo esatto $h(A_c) + h(A_t) = 0$ è garantita teoricamente, a patto che l'ottimizzazione interna (rispetto a $A_c, A_t, \varphi, \theta, \psi$) raggiunga un punto stazionario e che la sequenza $\mu \rightarrow +\infty$ sia gestita con sufficiente lentezza da consentire all'ottimizzazione interna di convergere a ciascun passo.

Il gradiente di $h(A)$ rispetto a A , necessario per l'ottimizzazione con discesa del gradiente, è:

$$\partial h(A)/\partial A = (e^{\{A \circ A\}})^T \odot (2A)$$

dove la trasposizione dell'esponenziale matriciale segue dalla regola di differenziazione della traccia. Questo gradiente è sempre ben definito poiché l'esponenziale matriciale è analitica e ovunque differenziabile.

A.13 Analisi degli equilibri dell'ELBO: collasso e degenerazione

L'ELBO causale di CV-JEPA ammette tre comportamenti degeneri che l'architettura deve prevenire attraverso un'oculata scelta dei pesi relativi dei termini.

Il primo comportamento degenero è il collasso del posteriore causale: se il peso di T_5 è troppo elevato rispetto ai termini di ricostruzione T_1 e T_2 , l'ottimizzazione spinge q_φ^s verso il prior $p(s)$, producendo fattori causali s costanti indipendentemente dall'input. In questo regime T_3 e T_4 si stabilizzano con $A_c \cdot s \approx \mu_c$ e $A_t \cdot s \approx \mu_t$ per qualche valore medio, perdendo qualsiasi contenuto informativo sui fattori generativi. Il rimedio architetturale è la strategia β -VAE con annealing del peso di T_5 : si parte con un peso molto piccolo e si aumenta progressivamente durante l'addestramento.

Il secondo comportamento degenero è la copia predittiva: se il peso di T_6 è eccessivamente grande, il predittore $q_\psi(t | z_c)$ impara a replicare $q_\varphi^t(t | x_t)$ con piccolo errore semplicemente codificando informazioni sufficienti nel contesto z_c , aggirando la necessità di una vera previsione causale. Il meccanismo di stop-gradient sul ramo target — che in CV-JEPA si applica bloccando il flusso del gradiente da T_6 verso i parametri dell'encoder del target enc_t — è il rimedio standard ereditato dall'architettura JEPA originale.

Il terzo comportamento degenero è il collasso del vincolo di aciclicità: se μ cresce troppo rapidamente nella sequenza del Lagrangiano aumentato, il termine quadratico h^2 domina l'ottimizzazione interna prima che le componenti di ricostruzione e regolarizzazione abbiano trovato la loro soluzione, producendo matrici A_c e A_t con tutti i pesi molto piccoli (struttura quasi-zero) che soddisfano $h \approx 0$ ma non portano alcuna informazione causale. Il rimedio è il protocollo di scheduling di μ sviluppato da Yu et al. per DAGNNH, adattato al contesto di CV-JEPA: si inizia l'addestramento senza il termine T_7 , si prelaiona i parametri φ , θ , ψ per alcune epoche, e si introduce progressivamente il vincolo causale soltanto quando l'ELBO delle prime sei componenti ha raggiunto un valore ragionevole.

A.14 La connessione con l'identificabilità nonlineare-ICA

La derivazione dell'ELBO causale trova la sua giustificazione statistica profonda nella teoria dell'identificabilità del Causal Representation Learning, la cui analisi è svolta nel Capitolo 2 della monografia attraverso il teorema iVAE di

Khemakhem et al. È utile qui riepilogare sinteticamente come le condizioni di identificabilità si traducano in vincoli sull'ELBO.

Il teorema iVAE afferma che, in un modello a variabili latenti con prior condizionato su una variabile ausiliaria u (nel caso di CV-JEPA, le azioni o gli interventi osservati), i fattori latenti sono identificabili a meno di trasformazioni element-wise se tre condizioni sono soddisfatte: (i) la log-densità del prior $\log p(s \mid u)$ appartiene alla famiglia esponenziale con statistiche sufficienti di s ; (ii) la matrice delle statistiche sufficienti è di rango pieno per ogni valore di u ; (iii) la funzione generativa g è iniettiva. In CV-JEPA la condizione (i) è soddisfatta per prior gaussiani condizionali $p(s \mid a) = \mathcal{N}(s; \mu(a), \text{diag}(\sigma^2(a)))$ con dipendenza dall'azione a ; la condizione (ii) richiede sufficiente variabilità delle azioni osservate durante l'addestramento; la condizione (iii) è garantita dall'invertibilità degli encoder.

Quando queste condizioni sono soddisfatte, la massimizzazione dell'ELBO causale non converge soltanto verso un minimo locale della perdita di addestramento, ma verso rappresentazioni causalmente identificabili nel senso del teorema: i fattori s stimati corrispondono ai fattori causali veri a meno di una trasformazione ammissibile. Questo è il risultato teorico che giustifica l'intera architettura CV-JEPA come strumento di Causal Representation Learning, non soltanto come variante del VAE con penalità strutturale.

Appendice B — Teorie Why3 per i moduli elementari dei sistemi JEPA

B.1 Il paradigma della verifica formale modulare

L'approccio alla verifica formale dei sistemi JEPA che questa appendice formalizza muove da un'osservazione di base analizzata nel Capitolo 19 della monografia: le reti neurali profonde nella loro generalità non sono traducibili in teorie del primo ordine verificabili da theorem prover, ma i moduli con struttura algebrica esplicita che le compongono lo sono, almeno localmente o in classi di input opportunamente ristrette. La strategia proposta da Araiza-Illan, Eder e Richards (2014) per i sistemi di controllo a blocchi Simulink diventa così, con i necessari adattamenti, un framework di verifica modulare per le componenti strutturalmente semplici di un'architettura JEPA.

Why3 è un sistema di verifica deduttiva sviluppato dall'INRIA che interfaccia con SMT solver multipli — CVC3, CVC4, Z3, Alt-Ergo — e con theorem prover interattivi come Coq e Isabelle/HOL, fungendo da intermediario che traduce *goal* di verifica in formato WhyML nel linguaggio nativo di ciascun solver. Il ciclo di verifica in Why3 prevede quattro fasi: (1) la specifica del

comportamento atteso del modulo come teoria WhyML con assiomi, predicati e *goal* da dimostrare; (2) la trasformazione automatica dei *goal* in VC (Verification Conditions) attraverso il calcolo di predicate transformer WP (weakest precondition) di Dijkstra; (3) l'invio delle VC a uno o più SMT solver; (4) la restituzione del risultato Valid, Unknown, oppure Invalid con un controesempio. Una prova Valid è una garanzia formale: la proprietà vale per tutti i valori di input nel dominio specificato dalla teoria, non soltanto su un campione di test.

Le teorie Why3 sviluppate in questa appendice coprono tre classi di moduli: le operazioni di normalizzazione (LayerNorm e RMSNorm), il meccanismo di attenzione scaled dot-product nella sua forma deterministica linearizzata, e il positional encoding sinusoidale. Per ciascun modulo si fornisce: la specifica formale della funzione input-output, la formulazione della proprietà da verificare come goal Why3, e una discussione della trattabilità della verifica per ciascun solver.

B.2 Teoria Why3 per LayerNorm

B.2.1 Specifica algebrica di LayerNorm

La normalizzazione a strato (Layer Normalization, Ba et al. 2016) calcola, per un vettore di input $x \in \mathbb{R}^d$, l'output:

$$\ln(x)_i = \gamma_i \cdot (x_i - \mu(x)) / (\sigma(x) + \epsilon) + \beta_i \quad \text{per } i = 1, \dots, d$$

dove:

$$\begin{aligned} \mu(x) &= (1/d) \sum_i x_i && \text{(media empirica dell'input)} \\ \sigma(x) &= \sqrt{(1/d) \sum_i (x_i - \mu(x))^2} && \text{(deviazione standard empirica)} \\ \epsilon &> 0 && \text{(costante di stabilità numerica, tipicamente } 1e-5 \text{ o } 1e-6) \\ \gamma &\in \mathbb{R}^d, \beta \in \mathbb{R}^d && \text{(parametri appresi di riscaldamento e traslazione)} \end{aligned}$$

Le proprietà di LayerNorm verificabili formalmente riguardano la stabilità numerica (il denominatore è sempre strettamente positivo?) e la correttezza del range dell'output (l'output normalizzato prima dell'applicazione di γ e β appartiene a un intervallo limitato?).

B.2.2 Proprietà da verificare: positività del denominatore

La proprietà cruciale è che il denominatore $\sigma(x) + \epsilon$ sia strettamente positivo per ogni $x \in \mathbb{R}^d$. Si verifica:

Proposizione B.1 (Positività del denominatore di LayerNorm):
 Per ogni $x \in \mathbb{R}^d$ e per ogni $\epsilon > 0$, vale $\sigma(x) + \epsilon > 0$.

La dimostrazione è elementare: $\sigma(x) = \sqrt{((1/d) \sum_i (x_i - \mu(x))^2)} \geq 0$ perché è la radice quadrata di una quantità non-negativa. Poiché $\epsilon > 0$, si ha $\sigma(x) + \epsilon \geq \epsilon > 0$. La proposizione è quindi vera per ogni $x \in \mathbb{R}^d$, inclusi i casi degeneri $x = c \cdot 1$ (vettore costante) in cui $\sigma(x) = 0$: il termine ϵ garantisce che la divisione sia sempre ben definita.

B.2.3 Formalizzazione Why3

La teoria Why3 per LayerNorm si struttura come segue. Si usa la logica del primo ordine con aritmetica reale, che è il frammento gestito efficacemente dagli SMT solver. Le variabili sono dichiarate come costanti simboliche nel contesto della teoria.

```
theory LayerNorm
  use real.Real
  use real.Square
  use real.Sqrt

  (* Costanti del modulo *)
  constant d : int      (* dimensionalità, d ≥ 1 *)
  axiom d_pos : d >= 1
  constant eps : real    (* ε > 0 *)
  axiom eps_pos : eps > 0.0

  (* Input: vettore x ∈ ℝᵈ, rappresentato come funzione int → real *)
  function x : int -> real

  (* Media e deviazione standard *)
  function mu_x : real = (1.0 / real_of_int d) * sum_x d x
  function sigma_sq_x : real = (1.0 / real_of_int d) * sum_sq_dev d x mu_x
  function sigma_x : real = sqrt sigma_sq_x

  (* Assioma: sigma_sq è non-negativa *)
  axiom sigma_sq_nonneg : sigma_sq_x >= 0.0

  (* Goal 1: Positività del denominatore *)
  goal denom_positive : sigma_x + eps > 0.0

  (* Goal 2: Il denominatore non è zero *)
  goal denom_nonzero : sigma_x + eps <> 0.0

end
```

La dimostrazione del Goal 1 segue direttamente dall'assioma `sigma_sq_nonneg` (che implica $\sigma^2 \geq 0$) e dall'assioma `eps_pos`: per l'aritmetica reale, $\sigma \geq 0$ e $\epsilon > 0$ implicano $\sigma + \epsilon \geq \epsilon > 0$. CVC3 e Z3 dimostrano entrambi questi goal istantaneamente (< 1 secondo) poiché si riducono a un'implicazione lineare nell'aritmetica reale.

B.2.4 Proprietà di equivarianza rispetto alle traslazioni

Un'importante proprietà strutturale di LayerNorm è la sua equivarianza rispetto a traslazioni costanti dell'input. Formalmente, per ogni $c \in \mathbb{R}$:

Proposizione B.2 (Equivarianza per traslazione):

$\text{ln}(x + c \cdot 1) = \text{ln}(x)$ (con uguaglianza in tutti i componenti)

La dimostrazione procede calcolando: $\mu(x + c \cdot 1) = \mu(x) + c$, quindi $x_i + c - \mu(x + c \cdot 1) = x_i + c - \mu(x) - c = x_i - \mu(x)$. La deviazione standard $\sigma(x + c \cdot 1) = \sigma(x)$ poiché dipende soltanto dagli scarti dalla media. Sostituendo nella formula di LayerNorm si ottiene l'identità.

Questa proprietà si traduce in Why3 come goal:

```
goal translation_equivariance :
  forall c : real.
    forall i : int. 1 <= i <= d ->
      layernorm (translate x c) i = layernorm x i
```

dove `translate x c` denota la funzione $\text{int} \rightarrow \text{real}$ definita da $(\text{translate } x \ c)(i) = x(i) + c$. La dimostrazione automatica di questo goal richiede che il solver possa ragionare sull'uguaglianza di due funzioni definite per composizione di operazioni aritmetiche — un tipo di ragionamento che CVC4 e Z3 gestiscono attraverso la teoria dei quantificatori sull'aritmetica reale, tipicamente in meno di 30 secondi per $d \leq 10$.

B.3 Teoria Why3 per RMSNorm

B.3.1 Specifica di RMSNorm

La normalizzazione per root-mean-square (RMSNorm, Zhang e Sennrich 2019) è una variante semplificata di LayerNorm che omette la sottrazione della media e calcola:

$\text{rms}(x)_i = \gamma_i \cdot x_i / (\text{rms_norm}(x) + \epsilon)$ per $i = 1, \dots, d$

dove:

$$\text{rms_norm}(x) = \sqrt{(1/d) \sum_i x_i^2}$$

L'omissione della sottrazione della media rende RMSNorm computazionalmente più efficiente e, crucialmente per la verifica formale, strutturalmente più semplice: la funzione `rms_norm(x)` è una norma euclidea riscalata, per la quale le proprietà di positività e non-nullità per input non-zero sono più semplici da specificare.

B.3.2 Proprietà verificabili per RMSNorm

La teoria Why3 per RMSNorm include i seguenti goal:

```
theory RMSNorm
  use real.Real
  use real.Square
  use real.Sqrt

  constant d : int
  axiom d_pos : d >= 1
  constant eps : real
  axiom eps_pos : eps > 0.0
  function x : int -> real

  function rms_sq : real = (1.0 / real_of_int d) * sum_squares d x
  function rms_val : real = sqrt rms_sq

  (* Assioma: rms_sq è non-negativa *)
  axiom rms_sq_nonneg : rms_sq >= 0.0

  (* Goal 1: Positività del denominatore *)
  goal denom_positive : rms_val + eps > 0.0

  (* Goal 2: Scala del denominatore rispetto alla norma dell'input *)
  (* rms_val = ||x||2 / √d — questa relazione è utile per bounds sull'output *)
  goal rms_l2_relation :
    rms_sq * real_of_int d = sum_squares d x

  (* Goal 3: Omogeneità — RMSNorm(λx) = sign(λ) · RMSNorm(x) per λ ≠ 0 *)
  (* Nota: la verifica richiede il trattamento del segno, fuori dalla logica lineare *)

end
```

Il Goal 2 è una proprietà algebrica elementare che tutti i solver dimostrano istantaneamente. Il Goal 3 sull'omogeneità introduce il segno, che richiede ragionamento per casi ($\lambda > 0$ oppure $\lambda < 0$) — un tipo di analisi che Why3 gestisce attraverso la tattica `split` che decompone il goal in sotto-goal per ciascun caso.

B.4 Teoria Why3 per il Positional Encoding sinusoidale

B.4.1 Specifica del Positional Encoding

Il positional encoding sinusoidale di Vaswani et al. assegna a ogni posizione $i \in \{0, 1, \dots, L-1\}$ e ogni dimensione $k \in \{0, 1, \dots, d/2-1\}$ la codifica:

$$\begin{aligned} \text{pe}(i, 2k) &= \sin(i / 10000^{2k/d}) \\ \text{pe}(i, 2k+1) &= \cos(i / 10000^{2k/d}) \end{aligned}$$

dove d è la dimensionalità del modello (assunta pari). Questo schema garantisce che la codifica di posizioni diverse sia distinguibile e che le relazioni di posizione relativa abbiano una struttura algebrica predicibile.

B.4.2 Proprietà verificabili: range e distinguibilità

Le proprietà di interesse per la verifica formale sono due. La prima è il range dell'output: le componenti del positional encoding appartengono sempre a $[-1, 1]$.

Proposizione B.3 (Range del positional encoding):

Per ogni $i \in \mathbb{Z}$ e ogni $k \in \{0, \dots, d/2-1\}$:

$$-1 \leq \text{pe}(i, 2k) \leq 1 \quad \text{e} \quad -1 \leq \text{pe}(i, 2k+1) \leq 1$$

La dimostrazione segue direttamente dalle proprietà standard delle funzioni seno e coseno: $\forall \theta \in \mathbb{R}, \sin(\theta) \in [-1, 1]$ e $\cos(\theta) \in [-1, 1]$.

La formalizzazione Why3 richiede la disponibilità degli assiomi standard per le funzioni trigonometriche nell'aritmetica reale. Why3 include la libreria `real.Trigonometry` che fornisce gli assiomi:

```
axiom sin_bound : forall theta : real. -1.0 <= sin theta <= 1.0
axiom cos_bound : forall theta : real. -1.0 <= cos theta <= 1.0
```

Con questi assiomi il goal di range si dimostra immediatamente:

```

theory PositionalEncoding
  use real.Real
  use real.Trigonometry

  constant d : int
  axiom d_pos : d >= 2
  axiom d_even : exists k. d = 2 * k

  constant base : real = 10000.0
  axiom base_pos : base > 0.0

  function freq (k : int) : real = 1.0 / pow base (2.0 * real_of_int k / real_of_int
d)

  function pe_sin (i : int) (k : int) : real = sin (real_of_int i * freq k)
  function pe_cos (i : int) (k : int) : real = cos (real_of_int i * freq k)

  (* Goal 1: range delle componenti sinusoidali *)
  goal sin_in_range :
    forall i k : int. -1.0 <= pe_sin i k <= 1.0

  goal cos_in_range :
    forall i k : int. -1.0 <= pe_cos i k <= 1.0

end

```

Il Goal 1 si dimostra per entrambe le componenti utilizzando direttamente gli assiomi di [real.Trigonometry](#). Z3 e Alt-Ergo dimostrano entrambi questi goal in meno di un secondo.

B.4.3 La proprietà di traslazione e il suo limite di verificabilità

Una proprietà più sofisticata del positional encoding sinusoidale è che, per ogni offset fisso δ , il positional encoding di posizione $i+\delta$ si esprime come trasformazione lineare del positional encoding di posizione i . Questa proprietà, che Pearl et al. (2023) usano per motivare l'uso del positional encoding sinusoidale nei transformer rispetto a encoding appresi, afferma:

$$\begin{aligned}
 \text{pe}(i+\delta, 2k) &= \text{pe}(i, 2k) \cdot \cos(\delta \cdot \text{freq}(k)) + \text{pe}(i, 2k+1) \cdot \sin(\delta \cdot \text{freq}(k)) \\
 \text{pe}(i+\delta, 2k+1) &= -\text{pe}(i, 2k) \cdot \sin(\delta \cdot \text{freq}(k)) + \text{pe}(i, 2k+1) \cdot \cos(\delta \cdot \text{freq}(k))
 \end{aligned}$$

Questa è semplicemente la formula dell'addizione del seno e del coseno applicata a $\theta = i \cdot \text{freq}(k)$ e $\varphi = \delta \cdot \text{freq}(k)$. La formalizzazione Why3 richiede l'assioma dell'addizione:

axiom sin_add : forall a b : real. sin (a + b) = sin a * cos b + cos a * sin b
 axiom cos_add : forall a b : real. cos (a + b) = cos a * cos b - sin a * sin b

che sono presenti nella libreria [real.Trigonometry](#) di Why3. Il goal di traslazione diventa:

```
goal translation_property :
  forall i delta k : int.
    pe_sin (i + delta) k =
      pe_sin i k * pe_cos_val delta k + pe_cos i k * pe_sin_val delta k
```

dove $pe_sin_val\ delta\ k = \sin(real_of_int\ delta * freq\ k)$ e $pe_cos_val\ delta\ k = \cos(real_of_int\ delta * freq\ k)$. Questo goal si riduce alla formula di addizione del seno e Z3 lo dimostra in pochi secondi, confermando che la struttura algebrica del positional encoding è verificabile formalmente in modo completo.

B.5 Teoria Why3 per il meccanismo di attenzione scaled dot-product (linearizzazione locale)

B.5.1 Il limite della verifica globale

Il meccanismo di attenzione scaled dot-product calcola:

$$Att(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

dove $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{n \times d_k}$, $V \in \mathbb{R}^{n \times d_v}$ sono le matrici di query, key e value. La presenza del softmax — una funzione non-lineare che normalizza ogni riga di $QK^T / \sqrt{d_k}$ sulla simplex (d_k-1) — rende impossibile la verifica globale delle proprietà del meccanismo di attenzione con SMT solver basati sull'aritmetica reale, poiché le equazioni risultanti sono trascendenti nel caso generale.

La strategia di verifica modulare prevede invece la verifica di proprietà della linearizzazione locale dell'attenzione attorno a un punto di operazione fissato. Data la jacobiana J_Att del meccanismo di attenzione calcolata in un punto (Q_0, K_0, V_0) , la linearizzazione locale afferma:

$$Att(Q_0 + \Delta Q, K_0 + \Delta K, V_0 + \Delta V) \approx Att(Q_0, K_0, V_0) + J_Att \cdot [\Delta Q; \Delta K; \Delta V] + O(\|\Delta\|^2)$$

La verifica della correttezza di questa approssimazione lineare all'ordine zero in $\|\Delta\|$ — cioè la verifica che J_Att sia calcolata correttamente — è un problema di algebra lineare che rientra nel dominio di verifica degli SMT solver.

B.5.2 Proprietà del softmax verificabili formalmente

Prima di procedere alla linearizzazione, si elencano le proprietà del softmax che ammettono formalizzazione Why3 diretta. Per un vettore $a \in \mathbb{R}^d$:

$$\text{softmax}(a)_i = \exp(a_i) / \sum \exp(a)$$

Le proprietà verificabili sono:

Proposizione B.4 (Proprietà del softmax):

- (i) $\text{softmax}(a)_i > 0$ per ogni i
- (ii) $\sum_i \text{softmax}(a)_i = 1$
- (iii) $\text{softmax}(a + c \cdot 1) = \text{softmax}(a)$ per ogni $c \in \mathbb{R}$ (invarianza per traslazione)

La proprietà (i) segue dalla positività di $\exp$; la (ii) è la proprietà di normalizzazione per costruzione; la (iii) segue dall'algebrismo $\exp(a_i + c) / \sum \exp(a + c) = \exp(c) \cdot \exp(a_i) / (\exp(c) \cdot \sum \exp(a)) = \exp(a_i) / \sum \exp(a)$.

La formalizzazione Why3 richiede la libreria `real.ExpLog` con gli assiomi standard per l'esponenziale:

```
theory Softmax
  use real.Real
  use real.ExpLog

  constant d : int
  axiom d_pos : d >= 1
  function a : int -> real

  function exp_sum : real = sum_exp d a

  axiom exp_sum_pos : exp_sum > 0.0 (* positività della somma di
  esponenziali *)

  function softmax_i (i : int) : real = exp (a i) / exp_sum

  (* Goal 1: positività *)
  goal softmax_positive :
    forall i : int. 1 <= i <= d -> softmax_i i > 0.0

  (* Goal 2: normalizzazione (verifica numerica approssimata per d finito) *)
```

```

goal softmax_sums_to_one :
  sum_softmax d softmax_i = 1.0

(* Goal 3: invarianza per traslazione *)
goal softmax_translation_invariant :
  forall c : real. forall i : int. 1 <= i <= d ->
    softmax_translated i c = softmax_i i

end

```

dove $\text{softmax_translated } i \ c$ denota $\text{softmax}(a + c \cdot 1)_i$. Il Goal 1 si dimostra dalla positività di $\exp$ e da $\exp_sum_pos$. Il Goal 2 richiede ragionamento su somme finite con cancellazioni algebriche, che CVC4 gestisce in modo automatico per $d \leq 4$ ma può richiedere assistenza interattiva (ad esempio attraverso Coq via Why3) per d arbitrario. Il Goal 3 si dimostra algebricamente con la regola $\exp(a_i + c) = \exp(a_i) \cdot \exp(c)$ e la distributività della divisione.

B.5.3 Linearizzazione locale e stabilità dell'attenzione

Per la linearizzazione locale del meccanismo di attenzione è conveniente separare le componenti. Si definisce la matrice dei logit $S \in \mathbb{R}^{n \times n}$ come:

$$S = QK^T / \sqrt{d_k}$$

La jacobiana del softmax applicata alla riga i della matrice S rispetto a S stessa è:

$$\partial \text{softmax}(S_i) / \partial S_i = \text{softmax}(S_i) \cdot (\delta - \text{softmax}(S_i))$$

dove δ è il delta di Kronecker. Questa jacobiana è la matrice diagonale della distribuzione di attenzione meno il suo prodotto esterno con se stessa, ed è simmetrica e semidefinita negativa. La jacobiana complessiva del meccanismo di attenzione rispetto a Q, K, V è:

$$\begin{aligned} \partial \text{Att} / \partial V &= \text{softmax}(QK^T / \sqrt{d_k}) \quad (\text{la matrice di attenzione, riga per riga}) \\ \partial \text{Att} / \partial Q &= (1 / \sqrt{d_k}) \cdot J_{\text{softmax}} \cdot K \cdot V^T \quad (\text{in forma semplificata per una testa}) \\ \partial \text{Att} / \partial K &= (1 / \sqrt{d_k}) \cdot J_{\text{softmax}} \cdot Q \cdot V^T \quad (\text{analogo}) \end{aligned}$$

La proprietà di stabilità verificabile formalmente è che la norma spettrale della jacobiana $\partial \text{Att} / \partial V$ sia limitata dalla norma dell'identità, poiché $\text{softmax}(\cdot)$ è una mappa sulla simplex con norma di righe uguale a 1:

$$\|\partial \text{Att} / \partial V\|_F \leq \sqrt{n} \quad (\text{norma di Frobenius della jacobiana rispetto a } V)$$

dove n è il numero di token. Questa stima segue dalla disuguaglianza di Cauchy-Schwarz applicata ai pesi di attenzione (che sommano a 1) e non richiede ragionamento non-lineare — è verificabile formalmente come conseguenza della normalizzazione del softmax.

B.6 Proprietà di equivarianza locale degli encoder JEPA

B.6.1 La proprietà di equivarianza come goal Why3

Il Capitolo 20 della monografia identifica la verifica dell'equivarianza degli encoder JEPA rispetto a gruppi di trasformazione finiti come terzo livello del programma di verifica formale modulare. La proprietà da verificare è:

$$\forall g \in G, \forall x \in \mathcal{X}: f_{\theta}(g \cdot x) = \varrho_g \cdot f_{\theta}(x)$$

dove G è un gruppo finito di trasformazioni (rotazioni di 90° , riflessioni), ϱ_g è la sua rappresentazione nello spazio latente, e f_{θ} è l'encoder.

Poiché f_{θ} è in generale non-lineare, la verifica globale è intrattabile. La strategia di verifica locale prevede invece di verificare la proprietà sulla jacobiana $J_{\{f_{\theta}\}}(x_0)$ calcolata in un punto specifico x_0 . La condizione di equivarianza infinitesimale è:

$$J_{\{f_{\theta}\}}(g \cdot x_0) \cdot Dg(x_0) = \varrho_g \cdot J_{\{f_{\theta}\}}(x_0)$$

dove $Dg(x_0)$ è la matrice jacobiana della trasformazione g nell'input space valutata in x_0 . Per trasformazioni lineari g (come le riflessioni e le rotazioni di 90° su immagini), $Dg(x_0) = Dg$ è una matrice costante indipendente da x_0 , il che semplifica drasticamente la condizione.

B.6.2 Formalizzazione per il gruppo C_4 (rotazioni discrete)

Per il gruppo ciclico C_4 delle rotazioni di 90° , con generatore g che manda la posizione (i, j) nella posizione $(j, H-1-i)$ su un'immagine di altezza H e larghezza W (con $H = W$ per semplicità), la matrice Dg è una matrice di permutazione $P_g \in \{0,1\}^{\{HW \times HW\}}$ con $P_g[P_g(i), i] = 1$ e 0 altrove.

La proprietà di equivarianza infinitesimale della jacobiana dell'encoder diventa:

$$J_{\{f_{\theta}\}}(P_g \cdot x_0) \cdot P_g = \varrho_g \cdot J_{\{f_{\theta}\}}(x_0)$$

che è un sistema di equazioni lineari nella jacobiana — verificabile formalmente, dato il valore numerico di $J_{\{f_\theta\}}(x_0)$ calcolato per un punto di operazione specifico.

La teoria Why3 corrispondente prevede:

```
theory LocalEquivariance
  use real.Real
  use matrix.Matrix

  (* Parametri del modulo *)
  constant d_in : int    (* dimensionalità input *)
  constant d_out : int   (* dimensionalità output *)
  axiom d_in_pos : d_in >= 1
  axiom d_out_pos : d_out >= 1

  (* Jacobiana dell'encoder nel punto  $x_0$  — trattata come costante simbolica *)
  function J : matrix real d_out d_in

  (* Matrice di permutazione della trasformazione g *)
  function Pg : matrix real d_in d_in

  (* Rappresentazione di g nello spazio latente *)
  function rho_g : matrix real d_out d_out

  (* Assiomi sulla struttura delle matrici di permutazione *)
  axiom Pg_permutation : is_permutation_matrix Pg
  axiom rho_g_orthogonal : mat_mul rho_g (transpose rho_g) = identity d_out

  (* Goal: equivarianza della jacobiana *)
  (* Si verifica su input  $x_0$  fisso attraverso equazioni lineari sulla jacobiana *)
  goal local_equivariance :
    mat_mul J_at_Pgx Pg = mat_mul rho_g J

end
```

dove $J_{\text{at_Pgx}}$ denota la jacobiana dell'encoder valutata in $P_g \cdot x_0$ — un'altra costante simbolica che il verificatore tratta come incognita vincolata dagli assiomi del modello. La verifica di questo goal richiede il ragionamento sull'algebra matriciale, che Z3 e CVC4 gestiscono attraverso la teoria delle matrici concrete (per dimensioni d_{in} , d_{out} piccole) o attraverso l'assiomatizzazione astratta.

B.7 Limiti strutturali della verifica Why3 per i sistemi JEPA

L'analisi condotta nelle sezioni precedenti permette di delineare con precisione i confini entro cui il framework Why3 è applicabile ai sistemi JEPA.

Il primo limite riguarda la dimensionalità: le teorie presentate nelle sezioni B.2-B.6 sono formulate per dimensioni d , n , d_{in} , d_{out} parametriche ma le dimostrazioni automatiche degli SMT solver diventano intrattabili per dimensioni superiori a un centinaio. Per encoder con spazi latenti di dimensione 256 o superiore — tipica per I-JEPA e V-JEPA — la verifica formale globale tramite SMT non è praticabile; la verifica deve essere condotta su sotto-moduli isolati (un singolo blocco di attenzione, una singola testa) con dimensioni ridotte.

Il secondo limite concerne la non-linearità: i moduli non-lineari — le funzioni di attivazione GELU e SiLU, la normalizzazione nella testa di attenzione, il softmax per sequenze lunghe — non rientrano nel frammento dell'aritmetica reale gestito dagli SMT solver senza approssimazioni. Le proprietà verificate nelle sezioni precedenti richiedono esplicitamente di isolare i moduli lineari (proiezioni) o di usare assiomi specifici (`sin_bound`, `cos_bound`, `exp_pos`) che il solver accetta senza dover derivarli autonomamente.

Il terzo limite, il più fondamentale, è l'assenza di specificazione consensuale per le proprietà semantiche di alto livello. La proprietà di stabilità numerica del denominatore di LayerNorm e la proprietà di range del positional encoding sono verificabili proprio perché hanno definizioni formali precise e univoche. Al contrario, la proprietà di «causally disentangled representation» o di «allineamento ai valori» non ha una definizione formale universalmente accettata nel Causal Representation Learning, come documentato nel Capitolo 2 della monografia. Senza una specificazione formale della proprietà da verificare, il problema della verifica formale è mal posto, non mal risolto.

Questi tre limiti non invalidano il programma di verifica formale modulare — lo delimitano con precisione. Le garanzie che il framework produce sono parziali ma non triviali: sapere che il denominatore di LayerNorm è sempre positivo, che il positional encoding è in $[-1, 1]$, e che la jacobiana dell'encoder in un punto specifico soddisfa l'equivarianza rispetto a rotazioni discrete è informazione di sicurezza reale, che i test empirici da soli non possono garantire perché il test copre sempre un insieme finito di input mentre la verifica formale copre il dominio intero (per le proprietà a cui si applica).

Appendice C — Prova del bound di Bernstein per lo stimatore misto nel framework delle etichette complementari

C.1 Contesto e motivazione

Il framework delle etichette complementari di Yin et al. per il scalable oversight senza ground truth introduce uno stimatore misto $\hat{A}_{\text{mix}}$ che combina linearmente lo stimatore ordinario $\hat{A}_{\text{ord}}$ e lo stimatore complementare $\hat{A}_{\text{comp}}$ al fine di minimizzare la varianza della stima dell'accuratezza A di un sistema LLM su task a K opzioni. La validità statistica di questo stimatore per campioni di dimensione finita — fondamentale in applicazioni di scalable oversight dove i dataset di valutazione hanno dimensione necessariamente limitata dalla scarsità di supervisor esperti — è garantita da un bound di deviazione di tipo Bernstein che sfrutta la struttura della varianza del mix in modo adattivo.

Questa appendice fornisce una derivazione completa del bound di Bernstein per $\hat{A}_{\text{mix}}$ enunciato nel Teorema 4 del paper di Yin et al., colmando il gap tra l'enunciato e la dimostrazione, che nel testo originale è affidata in parte a richiami della letteratura sui bound non-parametrici. La trattazione qui presentata è autonoma e fa riferimento alla disuguaglianza di Bernstein per variabili aleatorie sub-esponenziali nella forma di Boucheron, Lugosi e Massart (2013).

C.2 Notazione e definizioni del modello statistico

Si fissa un task a K opzioni con accuratezza vera $A \in (0, 1)$ e distribuzione complementare $q = (1 - A) / (K - 1)$. Si dispone di due insiemi di osservazioni indipendenti:

Osservazioni ordinarie: $X_1, \dots, X_{\{n_o\}} \sim \text{iid Bernoulli}(A)$

Osservazioni complementari: $Y_1, \dots, Y_{\{n_c\}} \sim \text{iid Bernoulli}(q)$

dove le $X_i \in \{0, 1\}$ indicano se il modello LLM risponde correttamente al quesito i , e le $Y_j \in \{0, 1\}$ indicano se il modello risponde correttamente alla domanda complementare j (ossia se la risposta del modello coincide con l'etichetta complementare, dove l'etichetta complementare è scelta uniformemente tra le $K-1$ risposte sbagliate). Le due sequenze sono indipendenti.

Si definiscono le statistiche sufficienti:

$$S_o = \sum_{i=1}^{n_o} X_i \sim \text{Binomiale}(n_o, A)$$

$$S_c = \sum_{i=1}^{n_c} Y_i \sim \text{Binomiale}(n_c, q)$$

da cui derivano i due stimatori di A:

$$\hat{A}_{ord} = S_o / n_o \quad (\text{stimatore ordinario})$$

$$\hat{A}_{comp} = (K-1) \cdot S_c / n_c - (K-2) \quad (\text{stimatore complementare})$$

È un fatto diretto che entrambi siano non distorti: $\mathbb{E}[\hat{A}_{ord}] = A$ e $\mathbb{E}[\hat{A}_{comp}] = (K-1)q - (K-2) = A$. Le varianze sono:

$$\text{Var}(\hat{A}_{ord}) = A(1-A) / n_o$$

$$\text{Var}(\hat{A}_{comp}) = (K-1)^2 \cdot q(1-q) / n_c$$

Lo stimatore misto con peso $w \in [0, 1]$ è:

$$\hat{A}_{mix}(w) = w \cdot \hat{A}_{ord} + (1-w) \cdot \hat{A}_{comp}$$

che è non distorto per qualsiasi w poiché entrambi i componenti lo sono. La varianza dello stimatore misto è:

$$\text{Var}(\hat{A}_{mix}(w)) = w^2 \cdot A(1-A)/n_o + (1-w)^2 \cdot (K-1)^2 \cdot q(1-q)/n_c$$

dove si è usata l'indipendenza di $\hat{A}_{ord}$ e $\hat{A}_{comp}$.

C.3 Il bound di Bernstein: enunciato preciso

Si enuncia qui il Teorema 4 di Yin et al. nella sua forma completa.

Teorema C.1 (Bound di Bernstein per $\hat{A}_{mix}$):

Per ogni $w \in [0, 1]$ e ogni $\delta \in (0, 1)$, con probabilità almeno $1 - \delta$:

$$|\hat{A}_{mix}(w) - A| \leq \sqrt{(2v \cdot \log(2/\delta))} + c \cdot \log(2/\delta)$$

dove:

$$v = w^2 \cdot A(1-A)/n_o + (1-w)^2 \cdot (K-1)^2 \cdot q(1-q)/n_c \quad (\text{varianza vera del mix})$$

$$c = \max(w/n_o, (1-w)(K-1)/n_c) \quad (\text{parametro di range})$$

Il bound è di tipo Bernstein nel senso che include un termine proporzionale a $\sqrt{v}$ e un termine proporzionale a c , dove v è la varianza vera dello stimatore e c è il massimo contributo di un singolo termine alla deviazione. La struttura in $\sqrt{v} + c \log(2/\delta)$ è migliore del puro bound di Hoeffding (che sarebbe proporzionale soltanto a $c\sqrt{\log(2/\delta)}$) quando la varianza è piccola, il che accade tipicamente quando q è vicino a 0 o 1 — cioè quando il modello valutato ha accuratezza A molto alta o molto bassa.

C.4 Strumenti probabilistici: la disuguaglianza di Bernstein per somme di variabili limitate

La derivazione si fonda sulla disuguaglianza di Bernstein per somme di variabili aleatorie indipendenti uniformemente limitate, nella formulazione seguente.

Lemma C.2 (Bernstein per variabili limitate):

Siano $Z_1, \dots, Z_n$ variabili aleatorie indipendenti con $\mathbb{E}[Z_i] = 0$, $|Z_i| \leq M$ quasi certamente,

e sia $\sigma^2 = (1/n) \sum_i \mathbb{E}[Z_i^2]$ la varianza media. Allora per ogni $t > 0$:

$$P(|\sum_i Z_i| \geq t) \leq 2 \cdot \exp(-t^2 / (2(n\sigma^2 + Mt/3)))$$

Applicando il Lemma C.2 alla somma centrata $\sum_i Z_i = \sum_i (Z_i - \mathbb{E}[Z_i])$ e invertendo la probabilità di fallimento $\delta = 2 \exp(-t^2 / (2(n\sigma^2 + Mt/3)))$, si ottiene il bound esplicito: con probabilità almeno $1 - \delta$:

$$|\sum_i Z_i / n| \leq \sqrt{(2\sigma^2 \log(2/\delta) / n)} + M \log(2/\delta) / (3n)$$

dove il fattore $1/3$ nel termine di range è assorbito nel coefficiente c della formulazione del Teorema C.1 (con costante leggermente diversa ma dello stesso ordine).

Una formulazione alternativa usa la disuguaglianza di Bernstein nella forma per la varianza totale, conveniente per lo stimatore misto:

$$P(|Z - \mathbb{E}[Z]| \geq t) \leq 2 \cdot \exp(-t^2 / (2(\text{Var}(Z) + M \cdot t / 3)))$$

dove $Z = \hat{A}_{\text{mix}}(w)$ è lo stimatore misto, $\text{Var}(Z) = v$ è la sua varianza, e M è il range massimo del singolo contributo.

C.5 Calcolo del range del contributo singolo in $\hat{A}_{\text{mix}}$

Lo stimatore misto si riscrive come somma di contributi indipendenti identificando il contributo di ogni osservazione singola. Definendo:

$W_i = w \cdot (X_i - A) / n_o$ per $i = 1, \dots, n_o$ (contributi dalle osservazioni ordinarie)

$V = (1-w) \cdot (K-1) \cdot (Y - q) / n_c$ per $j = 1, \dots, n_c$ (contributi dalle osservazioni complementari)

si ha:

$$\hat{A}_{\text{mix}}(w) - A = \sum_i W_i + \sum V$$

dove tutte le variabili $\{W_i\} \cup \{V\}$ sono centrate (media zero), indipendenti tra di loro, e con range uniformemente limitato. I range individuali sono:

$$|W_i| \leq w / n_o \quad \text{poiché } |X_i - A| \leq \max(A, 1-A) \leq 1$$

$$|V| \leq (1-w)(K-1) / n_c \quad \text{poiché } |Y - q| \leq \max(q, 1-q) \leq 1$$

Il range massimo del singolo termine — il M del Lemma C.2 — è pertanto:

$$M = \max(w/n_o, (1-w)(K-1)/n_c) = c$$

che è esattamente il parametro c del Teorema C.1. Questa identificazione mostra che c non è un parametro arbitrario ma il range effettivo del contributo dominante tra osservazioni ordinarie e complementari.

C.6 Calcolo della varianza totale di $\hat{A}_{\text{mix}} - A$

La varianza della variabile aleatoria $\hat{A}_{\text{mix}}(w) - A$ si ottiene sommando le varianze dei contributi indipendenti:

$$\text{Var}(\hat{A}_{\text{mix}}(w) - A) = \text{Var}(\sum_i W_i) + \text{Var}(\sum V)$$

Per i contributi ordinari:

$$\begin{aligned} \text{Var}(\sum_i W_i) &= \sum_i \text{Var}(W_i) = \sum_i w^2 \cdot \text{Var}(X_i) / n_o^2 \\ &= n_o \cdot w^2 \cdot A(1-A) / n_o^2 \end{aligned}$$

$$= w^2 \cdot A(1-A) / n_o$$

Per i contributi complementari:

$$\begin{aligned} \text{Var}(\sum V) &= \sum \text{Var}(V) = \sum (1-w)^2 \cdot (K-1)^2 \cdot \text{Var}(Y) / n_c^2 \\ &= n_c \cdot (1-w)^2 \cdot (K-1)^2 \cdot q(1-q) / n_c^2 \\ &= (1-w)^2 \cdot (K-1)^2 \cdot q(1-q) / n_c \end{aligned}$$

Sommando:

$$\text{Var}(\hat{A}_{\text{mix}}(w)) = w^2 \cdot A(1-A) / n_o + (1-w)^2 \cdot (K-1)^2 \cdot q(1-q) / n_c = v$$

che è esattamente il parametro di varianza v del Teorema C.1. Questa conferma algebrica mostra che la decomposizione del bound in termini di v è coerente con la struttura della varianza effettiva dello stimatore.

C.7 Applicazione del Lemma di Bernstein ad $\hat{A}_{\text{mix}} - A$

Con le definizioni stabilite nei paragrafi precedenti si può ora applicare il Lemma C.2 alla variabile aleatoria $Z = \hat{A}_{\text{mix}}(w) - A = \sum_i W_i + \sum V$, che è la somma di $n_o + n_c$ variabili aleatorie indipendenti centrate con range limitato $M = c$ e varianza totale $\text{Var}(Z) = v$.

La disuguaglianza di Bernstein nella forma per la varianza totale fornisce, per ogni $t > 0$:

$$P(|\hat{A}_{\text{mix}}(w) - A| \geq t) \leq 2 \cdot \exp(-t^2 / (2(v + c \cdot t / 3)))$$

Si vuole trovare t^* tale che il membro destro sia uguale a δ , ossia:

$$2 \cdot \exp(-t^{*2} / (2(v + c \cdot t^* / 3))) = \delta$$

prendendo i logaritmi:

$$t^{*2} / (2(v + c \cdot t^* / 3)) = \log(2/\delta)$$

che è un'equazione quadratica in t^* :

$$t^{*2} - 2 \cdot \log(2/\delta) \cdot c/3 \cdot t^* - 2v \cdot \log(2/\delta) = 0$$

Le soluzioni di questa quadratica sono:

$$t^* = \log(2/\delta) \cdot c/3 + \sqrt{(\log(2/\delta) \cdot c/3)^2 + 2v \cdot \log(2/\delta)}$$

Poiché si cerca il bound esplicito più semplice da comunicare, si migliora ulteriormente usando la disuguaglianza $\sqrt{a^2 + b} \leq a + \sqrt{b}$ (valida per $a, b \geq 0$):

$$\begin{aligned} t^* &\leq \log(2/\delta) \cdot c/3 + \log(2/\delta) \cdot c/3 + \sqrt{2v \cdot \log(2/\delta)} \\ &= 2 \cdot c \cdot \log(2/\delta) / 3 + \sqrt{2v \cdot \log(2/\delta)} \end{aligned}$$

Il coefficiente $2/3$ davanti a c è assorbito nel coefficiente c del Teorema C.1 (Yin et al. adottano la convenzione con il coefficiente 1 per semplicità espositiva, con la precisazione che il bound è una maggiorazione):

$$|\hat{A}_{\text{mix}}(w) - A| \leq \sqrt{2v \cdot \log(2/\delta)} + c \cdot \log(2/\delta) \quad \text{con probabilità } \geq 1 - \delta$$

che è esattamente l'enunciato del Teorema C.1. La derivazione è completa.

C.8 Ottimalità del bound per il peso IVW

Una conseguenza immediata del Teorema C.1 è che il bound di Bernstein per $\hat{A}_{\text{mix}}(w)$ è minimizzato, nella classe degli stimatori lineari, dalla scelta del peso IVW ottimo w^* . Il legame è diretto: poiché sia v che c dipendono da w , minimizzare v rispetto a w minimizza il termine dominante $\sqrt{2v \cdot \log(2/\delta)}$ del bound, e la minimizzazione della varianza v rispetto a w è un problema convesso con soluzione unica:

$$\begin{aligned} w^* &= \text{Var}(\hat{A}_{\text{comp}}) / (\text{Var}(\hat{A}_{\text{ord}}) + \text{Var}(\hat{A}_{\text{comp}})) \\ &= ((K-1)^2 \cdot q(1-q) / n_c) / (A(1-A)/n_o + (K-1)^2 \cdot q(1-q)/n_c) \end{aligned}$$

Con qualche algebra si ottiene la forma chiusa esplicita:

$$w^* = ((A + K - 2) \cdot n_o) / (A \cdot n_c + (A + K - 2) \cdot n_o)$$

che è la formula riportata nel Capitolo 16 della monografia. La varianza minima corrispondente è la media armonica delle varianze:

$$v_{\text{min}} = v(w^*) = \text{Var}(\hat{A}_{\text{ord}}) \cdot \text{Var}(\hat{A}_{\text{comp}}) / (\text{Var}(\hat{A}_{\text{ord}}) + \text{Var}(\hat{A}_{\text{comp}}))$$

Si verifica che $v(w^*) \leq v(w)$ per ogni $w \in [0, 1]$ sviluppando la differenza:

$$v(w) - v(w^*) = (w - w^*)^2 \cdot (\text{Var}(\hat{A}_{\text{ord}}) + \text{Var}(\hat{A}_{\text{comp}})) \geq 0$$

La disuguaglianza è ovviamente non-negativa poiché è un quadrato, con uguaglianza se e solo se $w = w^*$. Questo dimostra che w^* è il minimizzatore globale della varianza nella classe degli stimatori lineari non distorte, e che il bound di Bernstein con $w = w^*$ è il più stretto ottenibile in questa classe — enunciato riportato sinteticamente nel Capitolo 16 ma qui dimostrato completamente.

C.9 Il bound di Hoeffding come caso particolare

Per confronto si derivano qui brevemente il bound di Hoeffding corrispondente, che emerge dal limite $v \rightarrow 0$ (o equivalentemente dal caso in cui si ignori la struttura della varianza). La disuguaglianza di Hoeffding per somme di variabili limitate afferma: per variabili centrate Z_i con $|Z_i| \leq M$ quasi certamente:

$$P(|\sum_i Z_i| \geq t) \leq 2 \cdot \exp(-t^2 / (2n M^2))$$

Applicata ad $\hat{A}_{\text{mix}}(w) - A$ con $n = n_o + n_c$ termini e range $M = c$, il bound di Hoeffding produce:

$$|\hat{A}_{\text{mix}}(w) - A| \leq c \cdot \sqrt{2(n_o + n_c) \cdot \log(2/\delta)}$$

Questo bound è più semplice ma più conservativo del bound di Bernstein: cresce come $\sqrt{(n_o + n_c)}$ invece di $\sqrt{(v \cdot \log(1/\delta))}$. Il rapporto tra i due bound è approssimativamente:

$$\text{bound_Bernstein} / \text{bound_Hoeffding} \approx \sqrt{(v / (c^2 \cdot (n_o + n_c)))}$$

che è inferiore a 1 quando $v < c^2 \cdot (n_o + n_c)$, ossia quando la varianza effettiva del mix è più piccola del range quadratico scalato per la dimensione del campione. Poiché $v = O(1/n_o) + O(1/n_c)$ mentre $c^2 \cdot (n_o + n_c) = O(1)$, il bound di Bernstein è asintoticamente superiore al bound di Hoeffding per campioni grandi — una differenza che può essere di un fattore $\sqrt{n}$ per accuratezze estreme A vicino a 0 o 1.

C.10 Bound di Bernstein empirico: la variante adattiva

Il Teorema C.1 dipende dalla varianza vera v , che coinvolge le quantità ignote A e q . In pratica queste sono stimate dai dati, e si sostituisce v con lo stimatore plug-in:

$$\hat{v} = w^2 \cdot \hat{A}_{\text{ord}}(1 - \hat{A}_{\text{ord}})/n_o + (1-w)^2 \cdot (K-1)^2 \cdot \hat{q}(1-\hat{q})/n_c$$

dove $\hat{q} = S_c/n_c$. Il bound di Bernstein empirico risultante è:

$$|\hat{A}_{\text{mix}}(w) - A| \leq \sqrt{(2\hat{v} \cdot \log(4/\delta))} + c \cdot 7 \cdot \log(4/\delta) / 3 + c \cdot \epsilon_v$$

dove ϵ_v è l'errore di stima della varianza e il fattore $7/3$ proviene dalla maggiorazione dell'errore di plug-in. La derivazione di questa variante segue il schema di Maurer e Pontil (2009) per i bound empirici di Bernstein e non viene ripetuta qui nella sua interezza, ma si segnala che il fattore di correzione dalla varianza vera alla varianza empirica aggiunge soltanto un termine di ordine $O(\log(1/\delta)/n)$ rispetto al bound con varianza vera — trascurabile per n sufficientemente grande.

La forma mostrata nel Capitolo 16 della monografia per il bound dello stimatore complementare puro:

$$|\hat{A}_{\text{comp}} - A| \leq (K-1) \cdot \min\left(\sqrt{\log(2/\delta)} / (2 \cdot n_c), \sqrt{(2 \cdot \hat{q}(1-\hat{q}) \cdot \log(4/\delta))/(n_c-1)} + 7 \cdot \log(4/\delta)/(3 \cdot (n_c-1))\right)$$

è un caso particolare di questa struttura, dove il primo termine del min è il bound di Hoeffding e il secondo è il bound di Bernstein empirico applicato allo stimatore complementare $\hat{A}_{\text{comp}}$ — che è una funzione lineare della media campionaria S_c/n_c con coefficiente $(K-1)$, il che spiega la moltiplicazione per $(K-1)$ all'esterno del min.

C.11 Implicazioni per il protocollo di scalable oversight

Il bound di Bernstein per $\hat{A}_{\text{mix}}$ ha implicazioni pratiche dirette per il protocollo di valutazione senza ground truth nei sistemi di scalable oversight. La quantità di interesse per il supervisore è sapere se l'accuratezza A del sistema supera una soglia di sicurezza $A_{\text{threshold}}$ con incertezza controllata. La

domanda operativa è: quante osservazioni (n_o ordinarie e n_c complementari) sono necessarie per garantire un bound $|\hat{A}_{\text{mix}} - A| \leq \epsilon$ con probabilità $1 - \delta$?

Il Teorema C.1 risponde invertendo il bound: si ha $|\hat{A}_{\text{mix}}(w^*) - A| \leq \epsilon$ con probabilità almeno $1 - \delta$ se:

$$\sqrt{(2 v_{\min} \cdot \log(2/\delta))} + c_{\min} \cdot \log(2/\delta) \leq \epsilon$$

dove $v_{\min}$ e $c_{\min}$ si riferiscono al peso ottimo w^* . Per il regime tipico in cui il termine radice domina (ϵ grande rispetto a $c_{\min} \cdot \log(2/\delta)$), la condizione si semplifica in:

$$v_{\min} \leq \epsilon^2 / (2 \cdot \log(2/\delta))$$

Sostituendo la forma di $v_{\min} = 1 / (1/\text{Var}(\hat{A}_{\text{ord}}) + 1/\text{Var}(\hat{A}_{\text{comp}}))$:

$$1 / (n_o / (A(1-A)) + n_c / ((K-1)^2 \cdot q(1-q))) \leq \epsilon^2 / (2 \cdot \log(2/\delta))$$

da cui il numero totale di osservazioni richieste è:

$$n_o + n_c \geq 2 \cdot \log(2/\delta) \cdot (A(1-A) + (K-1)^2 \cdot q(1-q)) / \epsilon^2$$

nella decomposizione ottima tra n_o e n_c (che segue dalla minimizzazione della varianza totale a budget $n_o + n_c$ fissato). Il punto cruciale è che il budget di osservazioni cresce come $(K-1)^2 \cdot q(1-q) \approx (K-1)^2 / (K-1)^2 = 1/(K-1)$ per $A \approx 1/(K)$ (accuratezza casuale) — il che significa che per K grande il contributo delle etichette complementari domina e il numero di osservazioni complementari richieste scala come $1/(K-1)$, un risparmio lineare in K rispetto alla valutazione con sole etichette ordinarie.

Appendice D — Proposizioni ausiliarie: il do-calcolo, le condizioni iVAE di identificabilità, SIGReg e il teorema No Free Lunch per il Causal Representation Learning

D.1 Derivazione completa delle tre regole del do-calcolo di Pearl

Il do-calcolo è un sistema di tre regole di manipolazione formale delle probabilità interventali che, applicato iterativamente a un'espressione che coinvolge l'operatore $\text{do}(\cdot)$, permette di ridurre le quantità interventali a quantità osservative quando la struttura del DAG causale G soddisfa certe condizioni grafiche. La completezza del do-calcolo — dimostrata da Shpitser e Pearl nel 2006 — garantisce che ogni distribuzione interventale identificabile dal grafo G possa essere espressa in termini di quantità osservative applicando le tre regole un numero finito di volte.

Le tre regole operano su quattro insiemi di variabili X, Y, Z, W e su due grafi: il grafo originale G e i grafi mutilati $G_{\setminus X}$ (il grafo G con tutti gli archi entranti in X rimossi) e $G_{\setminus X}$ (il grafo G con tutti gli archi uscenti da X rimossi). La notazione $G_{\setminus XZ}$ denota il grafo con gli archi entranti in X e gli archi uscenti da Z simultaneamente rimossi.

Regola 1 (Inserzione/cancellazione di osservazioni):

$$P(y \mid \text{do}(x), z, w) = P(y \mid \text{do}(x), w) \quad \text{se } Y \perp_{\setminus G_{\setminus X}} Z \mid X, W$$

La condizione graficamente è che Z e Y siano d-separati da $X \cup W$ nel grafo $G_{\setminus X}$ — il grafo con gli archi entranti in X rimossi. Intuitivamente: se nell'esperimento in cui X è fissato per intervento, osservare Z non aggiunge informazione su Y rispetto a quanto già fornisce W , allora si può cancellare Z dalla condizionale.

La dimostrazione procede applicando il criterio di d-separazione al grafo $G_{\setminus X}$: la distribuzione $P(y \mid \text{do}(x), w)$ è la distribuzione marginale di Y nel modello mutilato $M_{\setminus x}$, e la d-separazione in $G_{\setminus X}$ garantisce che la distribuzione condizionale di Y dato X, Z, W nel modello mutilato fattorizzi eliminando la dipendenza da Z .

Regola 2 (Scambio azione-osservazione):

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), z, w) \quad \text{se } Y \perp_{\setminus G_{\setminus XZ}} Z \mid X, W$$

La condizione richiede che Z e Y siano d-separati in $G_{\setminus XZ}$ — il grafo con gli archi entranti in X e in Z entrambi rimossi — dato $X \cup W$. La regola afferma che quando questa condizione è soddisfatta, intervenire su Z è equivalente a osservare Z nella distribuzione manipolata da $\text{do}(x)$. In altre parole: Z è «come randomizzato» rispetto a Y nella distribuzione interventale su X .

La dimostrazione sfrutta il fatto che in $G_{\setminus XZ}$ la variabile Z non ha genitori causali attivi (sono stati rimossi dalla mutilazione), e quindi il meccanismo di Z nella distribuzione interventale $\text{do}(x)$ non differisce da quello della distribuzione osservativa dato X, W — da cui la sostituibilità tra $\text{do}(z)$ e z .

Regola 3 (Inserzione/cancellazione di azioni):

$$P(y \mid \text{do}(x), \text{do}(z), w) = P(y \mid \text{do}(x), w) \quad \text{se } Y \perp_{G_{\{X\bar{Z}(W)\}}} Z \mid X, W$$

dove $\bar{Z}(W)$ è il sottoinsieme di Z che non è antenato di alcun nodo di W in $G_{\{X\}}$. La condizione richiede la d-separazione nel grafo $G_{\{X\bar{Z}(W)\}}$, una variante del grafo mutilato in cui si rimuovono soltanto gli archi entranti nei nodi di Z che non sono antenati di W . L'intuizione è che se Z non ha effetto causale su Y dopo aver controllato per X e W — e dopo aver escluso i percorsi che passano per gli antenati di W — allora l'intervento $\text{do}(z)$ è irrilevante per $P(y)$.

Le tre regole sono corrette nel senso che ogni trasformazione da esse prodotta è valida per qualsiasi SCM con il DAG G dato. Sono complete nel senso che ogni quantità interventale identificabile può essere ridotta a quantità osservative applicandole iterativamente.

D.2 Il criterio del backdoor e la formula di aggiustamento

Tra le applicazioni del do-calcolo la formula di aggiustamento backdoor è quella di utilizzo più diretto per il Causal Representation Learning perché indica esplicitamente quali variabili devono essere condizionate per eliminare la confusione causale.

Definizione D.1 (Criterio del backdoor): Un insieme di variabili Z soddisfa il criterio del backdoor rispetto alla coppia (X, Y) nel DAG G se: (i) nessun elemento di Z è un discendente di X in G ; (ii) Z blocca ogni percorso backdoor tra X e Y , dove un percorso backdoor è un cammino non-orientato tra X e Y che inizia con un arco che punta verso X .

La formula di aggiustamento backdoor afferma: se Z soddisfa il criterio del backdoor rispetto a (X, Y) in G , allora:

$$P(Y = y \mid \text{do}(X = x)) = \sum_z P(Y = y \mid X = x, Z = z) \cdot P(Z = z)$$

ovvero la distribuzione interventale è ottenuta dalla distribuzione osservativa attraverso la standardizzazione (stratificazione) sulle variabili di confondimento Z .

La connessione con il framework CV-JEPA è strutturalmente diretta: se i fattori causali s identificati dall'encoder JEPA soddisfano il criterio del backdoor rispetto alla coppia (z_c, z_t) nello spazio latente — cioè se s blocca tutti i

percorsi backdoor tra il contesto e il target nello spazio latente — allora la formula di aggiustamento si applica e la distribuzione interventale $P(z_t \mid \text{do}(z_c))$ è calcolabile come:

$$P(z_t \mid \text{do}(z_c)) = \int P(z_t \mid z_c, s) \cdot p(s) \, ds$$

che è esattamente l'operazione computata dal predittore s_ψ di CV-JEPA quando i fattori causali s sono stati identificati dall'encoder. Questa connessione, sviluppata nel Capitolo 3 della monografia, mostra che il predittore JEPA non è soltanto un meccanismo di previsione ma, in condizioni di identificabilità, un meccanismo di calcolo delle distribuzioni interventali.

D.3 Le condizioni di identificabilità iVAE e la loro traduzione in vincoli architetturali

Il teorema fondamentale di identificabilità del Causal Representation Learning a noi rilevante è il risultato di Khemakhem et al. (2020) sul modello iVAE (Identifiable Variational Autoencoder). Nell'originale, il modello iVAE considera un generatore $g: \mathbb{R}^m \rightarrow \mathbb{R}^d$ con $m < d$ e un prior condizionato su una variabile ausiliaria u :

$$p(z \mid u) = \prod_i p_i(z_i \mid u) \quad (\text{prior fattorizzato condizionato su } u)$$

dove ciascun $p_i(z_i \mid u)$ appartiene alla famiglia esponenziale con statistiche sufficienti $T_i(z_i)$:

$$\log p_i(z_i \mid u) = T_i(z_i)^T \cdot \lambda_i(u) - \log Z_i(u)$$

con $\lambda_i(u)$ il parametro naturale e $Z_i(u)$ la funzione di partizione.

Teorema D.2 (Identificabilità di iVAE, Khemakhem et al. 2020): Sotto le seguenti tre condizioni, i fattori latenti z sono identificabili a meno di permutazione e trasformazioni element-wise:

(C1) Il generatore g è iniettivo e la sua jacobiana ha rango m quasi ovunque.

(C2) La log-densità del prior condizionale appartiene alla famiglia esponenziale con $k \geq 2$ statistiche sufficienti per ogni componente: $\log p_i(z_i \mid u) = \sum_{k \in \{1, \dots, k\}} T_{i,k}(z_i) \cdot \lambda_{i,k}(u) - \log Z_i(u)$.

(C3) Esistono $n+1$ valori $u^0, u^1, \dots, u^n$ della variabile ausiliaria tali che la matrice $L \in \mathbb{R}^{(kn) \times n}$ con elementi $L_{\{(i-1)k+l, j\}} = \lambda_{\{i,l\}}(u^j) - \lambda_{\{i,l\}}(u^0)$ abbia rango n (rango pieno per colonne).

Interpretazione delle condizioni nel contesto CV-JEPA. La variabile ausiliaria u in iVAE corrisponde, in CV-JEPA, alle azioni esterne a osservate durante la raccolta dei dati di addestramento. La condizione (C1) richiede che l'encoder enc_c sia iniettivo — condizione soddisfatta per encoder ViT a capacità sufficiente in regime non sovrapposto. La condizione (C2) richiede che il prior $p(s \mid a)$ sia della famiglia esponenziale con almeno due statistiche sufficienti per componente — soddisfatta per prior gaussiani con media e varianza dipendenti dall'azione: $p(s_i \mid a) = \mathcal{N}(\mu_i(a), \sigma_i^2(a))$, che ha statistiche sufficienti $T_1(s_i) = s_i$ e $T_2(s_i) = s_i^2$ con parametri naturali $\lambda_{1i}(a) = \mu_i(a)/\sigma_i^2(a)$ e $\lambda_{2i}(a) = -1/(2\sigma_i^2(a))$. La condizione (C3) richiede sufficiente variabilità delle azioni a : le azioni devono generare distribuzioni $p(s \mid a)$ sufficientemente diverse tra di loro, in modo che la matrice L abbia rango pieno. Questa condizione è una condizione sul dataset di addestramento, non sull'architettura: richiede che il training coverage copra azioni sufficientemente diverse.

D.4 SIGReg come implementazione del Principio dei Meccanismi Causali Indipendenti

Il Sketched Isotropic Gaussian Regularizer (SIGReg), introdotto da Balestriero e LeCun come strumento di regolarizzazione per gli spazi latenti JEPA, trova nel contesto causale una giustificazione che trascende la sua origine come penalità statistica. Il Principio dei Meccanismi Causali Indipendenti (Independent Causal Mechanisms, ICM) di Peters, Janzing e Schölkopf afferma che in un sistema causale genuino, i meccanismi che generano le diverse variabili endogene sono autonomi: la distribuzione condizionale $p(V_i \mid \text{Pa}(V_i))$ di ciascuna variabile non contiene informazione sulle distribuzioni condizionali delle altre variabili, e non viene alterata da interventi su variabili diverse da V_i stessa.

La SIGReg è definita formalmente come:

$$\text{SIGReg}(Z) \triangleq (1/M) \sum_{m=1}^M T(Z \cdot u^{\{(m)\}})$$

dove $u^{\{(1)\}}, \dots, u^{\{(M)\}}$ sono direzioni campionate uniformemente dalla sfera unitaria S^{d-1} , e $T(\cdot)$ è una statistica di test di normalità univariata — tipicamente il test di Cramér-von Mises o il test KS contro la distribuzione gaussiana standard $\mathcal{N}(0, 1)$. Per il teorema di Cramér-Wold: $\text{SIGReg}(Z) \rightarrow 0$ se e solo se la distribuzione marginale degli embedding P_Z converge alla

gaussiana isotropa $\mathcal{N}(0, I_d)$ nella metrica della distanza totale di variazione su proiezioni univariate.

La connessione con l'ICM procede nel modo seguente. Se i fattori latenti $s = (s_1, \dots, s_k)$ sono causalmente indipendenti nel senso del Principio ICM — ossia i meccanismi che li generano sono autonomi — allora nel frame di riferimento naturale (identificato dal DAG causale), i fattori $s_1, \dots, s_k$ devono essere stocasticamente indipendenti nella distribuzione marginale. Per il teorema di Darmois (1953), l'indipendenza marginale di variabili con distribuzione comune equivale alla gaussianità quando la trasformazione che le genera è lineare e invertibile — ma questo risultato non si estende al caso non-lineare. Tuttavia, imporre che la distribuzione marginale di Z sia gaussiana isotropa attraverso SIGReg è una condizione necessaria (non sufficiente) per l'indipendenza dei fattori: una distribuzione gaussiana isotropa $\mathcal{N}(0, I_d)$ ha componenti scorrelate, e per gaussiane multivariate la scorrelazione implica l'indipendenza. SIGReg impone dunque una forma di indipendenza di primo e secondo ordine tra i fattori — più debole dell'indipendenza piena richiesta dall'ICM ma computazionalmente trattabile e differenziabile rispetto ai parametri dell'encoder.

La forma kernel di SIGReg, discussa nel Capitolo 17 della monografia, si ottiene scegliendo T come maximum mean discrepancy (MMD) tra la distribuzione empirica delle proiezioni $Z \cdot u^{\wedge(m)}$ e la distribuzione gaussiana standard target. Con kernel gaussiano $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$:

$$T_{\text{MMD}}(Z \cdot u) = \mathbb{E}_{\{x, x' \sim P_{\{Z \cdot u\}}\}}[k(x, x')] - 2 \cdot \mathbb{E}_{\{x \sim P_{\{Z \cdot u\}}, y \sim \mathcal{N}(0,1)\}}[k(x, y)] + \mathbb{E}_{\{y, y' \sim \mathcal{N}(0,1)\}}[k(y, y')]$$

che è calcolabile in $O(n^2 d)$ per n embedding di dimensione d . La convergenza di $\text{SIGReg} \rightarrow 0$ è equivalente alla convergenza in MMD della distribuzione degli embedding verso la gaussiana target.

D.5 Il teorema No Free Lunch per il Causal Representation Learning

Il teorema No Free Lunch (NFL) nel dominio del Causal Representation Learning afferma, in termini generali, che nessun algoritmo di apprendimento delle rappresentazioni causali può avere prestazioni di identificabilità uniformemente superiori a tutti gli altri algoritmi su tutte le possibili distribuzioni dei dati e tutte le possibili strutture causali. La formulazione precisa nel contesto dei sistemi JEPa è la seguente.

Proposizione D.3 (NFL per il CRL): Sia $\mathcal{A}$ un algoritmo di Causal Representation Learning che, dato un dataset D di n osservazioni, produce una

rappresentazione latente $\hat{z}(D)$ che stima i fattori causali veri z . Per ogni misura di errore di identificabilità $\varepsilon(\hat{z}, z) \geq 0$, l'errore medio di $\mathcal{A}$ su tutte le distribuzioni dei dati compatibili con strutture causali di cardinalità fissata non dipende dall'algoritmo:

$$\mathbb{E}_{\{P \sim \mathcal{P}, D \sim P^n\}} [\varepsilon(\hat{z}_{\mathcal{A}}(D), z)] = \text{costante indipendente da } \mathcal{A}$$

dove $\mathcal{P}$ è la classe di tutte le distribuzioni compatibili con qualche struttura causale a m fattori.

Corollario D.4 (Non-universalità degli algoritmi CRL): Per ogni algoritmo $\mathcal{A}$ esiste una struttura causale $P^* \in \mathcal{P}$ su cui $\mathcal{A}$ ha errore di identificabilità pari a quello di un algoritmo casuale.

La dimostrazione della Proposizione D.3 segue la struttura standard del teorema NFL di Wolpert (1996) per l'apprendimento supervisionato, adattata al contesto del CRL. Si costruiscono due insiemi di distribuzioni: un sottoinsieme $\mathcal{P}_{\text{easy}} \subset \mathcal{P}$ su cui $\mathcal{A}$ performa bene (ipotesi di induzione causale corretta) e un sottoinsieme $\mathcal{P}_{\text{hard}} \subset \mathcal{P}$ su cui le stesse ipotesi portano $\mathcal{A}$ fuori strada. Il NFL afferma che $|\mathcal{P}_{\text{easy}}| = |\mathcal{P}_{\text{hard}}|$ — i due sottoinsiemi hanno la stessa cardinalità — quando si usa la misura uniforme su $\mathcal{P}$.

La costruzione esplicita di $\mathcal{P}_{\text{hard}}$ per un algoritmo basato sull'induzione causale verso la gaussianità (come SIGReg o iVAE) procede come segue. Si prenda una qualsiasi distribuzione $P \in \mathcal{P}_{\text{easy}}$ per cui l'algoritmo identifichi correttamente i fattori causali s . Si definisce P' come la distribuzione ottenuta da P attraverso una permutazione dei fattori: $s'_i = s_{\{\pi(i)\}}$ per qualche permutazione π . La distribuzione P' ha gli stessi fattori marginali di P ma con relazioni causali riordinate. Poiché la statistica SIGReg e il criterio iVAE sono invarianti per permutazione dei fattori (il prior gaussiano isotropo è simmetrico per permutazione), l'algoritmo $\mathcal{A}$ su P' produce la stessa rappresentazione $\hat{z}_{\mathcal{A}}$ di P — ma questa rappresentazione ora ha errore di identificabilità positivo perché associa al fattore $\pi(i)$ -esimo il ruolo causale del fattore i -esimo. L'insieme di queste distribuzioni permutate costituisce $\mathcal{P}_{\text{hard}}$, e per ogni $P \in \mathcal{P}_{\text{easy}}$ esiste esattamente un $P' \in \mathcal{P}_{\text{hard}}$ con struttura isomorfa. Dunque i due insiemi hanno la stessa cardinalità, e l'errore medio su $\mathcal{P}_{\text{easy}} \cup \mathcal{P}_{\text{hard}}$ è indipendente dall'algoritmo.

Conseguenza architettonica per CV-JEPA. Il NFL non implica che CV-JEPA sia inutile — implica che CV-JEPA performa meglio degli algoritmi alternativi soltanto sulle distribuzioni per cui la sua ipotesi causale (indipendenza dei fattori, struttura DAG aciclica, connessione lineare tra fattori causali e vista-specifici) è circa corretta. Quando questa ipotesi viene sistematicamente violata — ad esempio in domini con causalità ciclica o con connessioni altamente non-lineari tra fattori — il vantaggio di CV-JEPA rispetto ad algoritmi non causali scompare. Questo non è un difetto dell'approccio ma una conseguenza strutturale del NFL: il bias causale

introdotto da CV-JEPA è un vantaggio nei problemi per cui è appropriato e un onere nei problemi per cui non lo è.

D.6 Prova della convergenza del peso IVW plug-in

Il peso IVW plug-in $\hat{w}_{IVW}$ è ottenuto sostituendo A con $\hat{A}_{ord}$ nella formula del peso ottimo:

$$\hat{w}_{IVW} = \text{Var}(\hat{A}_{comp}) / (\text{Var}(\hat{A}_{ord}) + \text{Var}(\hat{A}_{comp}))$$

dove $\text{Var}(\hat{A}_{ord}) = \hat{A}_{ord}(1 - \hat{A}_{ord})/n_o$ e $\text{Var}(\hat{A}_{comp}) = (K-1)^2 \cdot \hat{q}(1-\hat{q})/n_c$ con $\hat{q} = \hat{A}_{comp}/(K-1) + (K-2)/(K-1)$. La proposizione che segue garantisce che l'errore introdotto dall'uso del peso plug-in anziché del peso ottimo vero sia trascurabile per campioni sufficientemente grandi.

Proposizione D.5 (Convergenza del peso IVW plug-in): Sotto le condizioni del modello statistico del paragrafo C.2, si ha:

$$|\text{Var}(\hat{A}_{mix}(\hat{w}_{IVW})) - \text{Var}(\hat{A}_{mix}(w^*))| = O_p(n_o^{-1} + n_c^{-1})$$

ossia l'eccesso di varianza introdotto dall'approssimazione del peso è dell'ordine n^{-1} in entrambi i campioni.

Dimostrazione. Si scrive il peso ottimo vero come $w^* = f(A, q, n_o, n_c)$ e il peso plug-in come $\hat{w} = f(\hat{A}_{ord}, \hat{q}, n_o, n_c)$. Per la differenziabilità di f in tutti gli argomenti, si espande in serie di Taylor al primo ordine:

$$\hat{w} - w^* = \partial f / \partial A \cdot (\hat{A}_{ord} - A) + \partial f / \partial q \cdot (\hat{q} - q) + O_p((\hat{A}_{ord} - A)^2 + (\hat{q} - q)^2)$$

Poiché $\hat{A}_{ord} - A = O_p(n_o^{-1/2})$ e $\hat{q} - q = O_p(n_c^{-1/2})$, si ha $\hat{w} - w^* = O_p(n_o^{-1/2} + n_c^{-1/2})$.

La varianza del mix è una funzione quadratica di w : $\text{Var}(\hat{A}_{mix}(w)) = w^2 \cdot \sigma_o^2 + (1-w)^2 \cdot \sigma_c^2$ con $\sigma_o^2 = A(1-A)/n_o$ e $\sigma_c^2 = (K-1)^2 q(1-q)/n_c$. Per il teorema della funzione composta:

$$\begin{aligned} |\text{Var}(\hat{A}_{mix}(\hat{w})) - \text{Var}(\hat{A}_{mix}(w^*))| &\leq |\partial \text{Var} / \partial w|_{w=w^*} \cdot |\hat{w} - w^*| + O_p((\hat{w} - w^*)^2) \\ &= O_p(n_o^{-1/2} + n_c^{-1/2}) \end{aligned}$$

dove si è usato il fatto che $|\partial \text{Var} / \partial w|_{w=w^*} = 0$ (perché w^* è un minimo della varianza e il gradiente si annulla in un minimo). L'espansione al secondo ordine produce:

$$|\text{Var}(\hat{A}_{\text{mix}}(\hat{w})) - \text{Var}(\hat{A}_{\text{mix}}(w^*))| = O_p((\hat{w} - w^*)^2) = O_p(n_o^{-1} + n_c^{-1})$$

che è il risultato enunciato. ■

Questo risultato implica che l'eccesso di varianza introdotto dall'approssimazione del peso è un termine di ordine n^{-1} , trascurabile rispetto alla varianza principale $O(n^{-1/2})$ dello stimatore, e che il bound di Bernstein del Teorema C.1 applicato con il peso plug-in $\hat{w}_{\text{IVW}}$ è asintoticamente equivalente al bound con il peso ottimo vero w^* a meno di termini di ordine $O(\log(1/\delta)/n)$.

Bibliografia

La presente raccolta bibliografica raccoglie i riferimenti citati o discussi in modo esplicito nel corpo della monografia e nelle appendici tecniche. L'ordinamento è alfabetico per cognome del primo autore. I paper senza anno di pubblicazione consolidato al momento della stesura sono indicati con l'anno del preprint arXiv laddove disponibile. Le opere enciclopediche e i libri di testo citati come fonti primarie della formalizzazione adottata sono distinti dalle voci di articoli scientifici attraverso l'indicazione del tipo di contributo tra parentesi quadre.

Araiza-Illan, D., Eder, K. e Richards, A. (2014). Formal Verification of Control Systems' Properties with Theorem Proving. In *Proceedings of the UKACC International Conference on Control*. Loughborough, UK. [Articolo in atti di conferenza]

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y. e Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [Articolo — proposta originale di I-JEPA]

Ba, J. L., Kiros, J. R. e Hinton, G. E. (2016). Layer Normalization. *arXiv preprint arXiv:1607.06450*. [Preprint]

Balestrieri, R. e LeCun, Y. (2023). Contrastive and Non-Contrastive Self-Supervised Learning Recover Global and Local Spectral Embedding Methods. In *Advances in Neural Information Processing Systems (NeurIPS)*. [Articolo — analisi teorica degli spazi JEPA e introduzione di LeJEPA]

Boucheron, S., Lugosi, G. e Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. [Libro di testo — fonte delle disuguaglianze di Bernstein e Hoeffding utilizzate in Appendice C]

Brooks, E., Walls, A., Zhou, Y., Bhardwaj, S., Doshi, A., Kossmann, D., Beal, J., Karabolas, N. e Bohg, J. (2024). Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. [Articolo — dataset e modelli della famiglia RT-X]

Chen, Z., Deng, Y., Yuan, H., Ji, K. e Gu, Q. (2022). Towards Deepening Graph Neural Networks: A GNTK-Based Optimization Perspective. In *International Conference on Learning Representations (ICLR)*. [Articolo — citato nel contesto del probing causale]

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21(1–2), 2–8. [Articolo classico — teorema di Darmois-Skitovic citato in Appendice D e nel Capitolo 2]

Dreamer (famiglia): si vedano le voci Hafner (2018, 2020, 2023) sotto.

Gögl, L. e Yau, S.-T. (2023). Var-JEPA: A Probabilistic Framework for Joint-Embedding Predictive Architectures. *arXiv preprint* [disponibile in preprint; la versione specifica del paper è quella citata nella seconda monografia del trittico]. [Preprint — fonte primaria della derivazione dell'ELBO a cinque termini di Var-JEPA e della connessione tra JEPA e i modelli generativi variazionali]

Hafner, D., Lillicrap, T., Ba, J. e Norouzi, M. (2018). Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the International Conference on Machine Learning (ICML)* (2019). *arXiv:1811.04551*. [Articolo — proposta originale dell'RSSM e del modello PlaNet]

Hafner, D., Lillicrap, T., Norouzi, M. e Ba, J. (2020). Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations (ICLR)*. [Articolo — DreamerV1]

Hafner, D., Lillicrap, T., Ba, J. e Norouzi, M. (2021). Mastering Atari with Discrete World Models. In *International Conference on Learning Representations (ICLR)*. [Articolo — DreamerV2]

Hafner, D., Pasukonis, J., Ba, J. e Lillicrap, T. (2023). Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*. [Preprint — DreamerV3; fonte della formalizzazione dell'RSSM e del confronto con l'ELBO di Var-JEPA nel Capitolo 11]

Hestenes, M. R. (1969). Multiplier and Gradient Methods. *Journal of Optimization Theory and Applications*, 4(5), 303–320. [Articolo classico — fonte del metodo del Lagrangiano aumentato citato in Appendice A]

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301), 13–30. [Articolo classico — bound di Hoeffding citato in Appendice C]

Huang, A., LeCun, Y. e Balestriero, R. (2023). LLM-JEPA: A Joint-Embedding Predictive Architecture for Language. *arXiv preprint* [citato nella seconda monografia del trittico]. [Preprint — estensione del framework JEPA ai modelli linguistici di grande scala]

Khemakhem, I., Kingma, D., Monti, R. e Hyvarinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

PMLR. [Articolo — proposta di iVAE e teorema di identificabilità; fonte primaria del Capitolo 2 e dell'Appendice D]

Kim, B. e Gilmer, J. (2018). TCAV: Quantitative Testing with Concept Activation Vectors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR. [Articolo — tecnica TCAV per l'interpretabilità, analizzata nel Capitolo 13]

Kingma, D. P. e Welling, M. (2014). Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. [Articolo — VAE originale, incluso il reparametrization trick citato in Appendice A]

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. *OpenReview preprint*. [Preprint — framework AMI con i sei moduli; fonte primaria dell'architettura di riferimento per la terza monografia]

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B. e Wu, J. (2019). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. [Articolo — NS-CL; analizzato nel Capitolo 4]

Mathieu, E., Rainforth, T., Siddharth, N. e Teh, Y. W. (2019). Disentangled Sequential Autoencoder. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR. [Articolo — ELBO surgery; la notazione compatta dell'ELBO causale in Appendice A segue questo lavoro]

Maurer, A. e Pontil, M. (2009). Empirical Bernstein Bounds and Sample Variance Penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*. [Articolo — bound di Bernstein empirico con varianza stimata; citato nella discussione dell'Appendice C]

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. [Libro di testo — fonte primaria del formalismo SCM, del do-calcolo, e della gerarchia causale]

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2a edizione). Cambridge University Press. [Libro di testo — edizione aggiornata; i controesempi canonici dell'irriducibilità causale del Capitolo 1 sono tratti dal Capitolo 3 di questa edizione]

Peters, J., Janzing, D. e Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press. [Libro di testo — Principio dei Meccanismi Causali Indipendenti (ICM) citato in Appendice D; testo di riferimento per l'intera Parte Prima della monografia]

Powell, M. J. D. (1969). A Method for Nonlinear Constraints in Minimization Problems. In R. Fletcher (a cura di), *Optimization*. Academic Press. [Capitolo in

volume — metodo del Lagrangiano aumentato, citato in parallelo con Hestenes (1969)]

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J. e Bowman, S. R. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*. [Preprint — benchmark GPQA citato nella validazione empirica del Capitolo 16]

Schmidhuber, J. (1991). A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers. In *From Animals to Animats: Proceedings of the First International Conference on the Simulation of Adaptive Behavior*. MIT Press. [Articolo classico — prime idee di World Model per la pianificazione]

Shpitser, I. e Pearl, J. (2006). Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*. [Articolo — completezza del do-calcolo; citato nel Capitolo 1 e in Appendice D]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. e Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30. [Articolo — architettura Transformer originale; fonte del positional encoding sinusoidale formalizzato in Appendice B]

Wen, T., Zhu, B., He, X., Nie, P., Wan, Y., Deng, J., Guo, L., Xu, Z. e Yuan, Z. (2025). Recursive Critique and Refinement for Scalable Oversight of Foundation Models. *arXiv preprint* [anno approssimativo basato sul contesto della monografia]. [Preprint — framework della recursive critique per il scalable oversight; analizzato nel Capitolo 16]

Wolpert, D. H. (1996). The Lack of A Priori Distinctions between Learning Algorithms. *Neural Computation*, 8(7), 1341–1390. [Articolo classico — teorema No Free Lunch per l'apprendimento; citato in Appendice D]

Wolpert, D. H. e Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. [Articolo classico — teorema No Free Lunch per l'ottimizzazione; citato nel Capitolo 19]

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J. e Wang, J. (2021). CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [Articolo — CausalVAE; analizzato nel Capitolo 2 e nelle appendici A e D]

Yin, Z. e collaboratori (2024–2025). Complementary Labels for Accuracy Estimation without Ground Truth. *arXiv preprint* [anno e coautori dedotti dal contesto del Capitolo 16; la versione specifica con i Teoremi 2 e 4 citati in

Appendice C è quella analizzata nel Capitolo 16]. [Preprint — framework delle etichette complementari per il scalable oversight; fonte primaria dell'Appendice C]

Yu, X., Chen, G., Yu, Y., Dong, Y., Zhu, Y., Liu, J., Tang, L., Zhang, Q., Tong, Y., Li, X. e Zhang, K. (2021). DAG-GNN: DAG Structure Learning with Graph Neural Networks. Estensione e aggiornamento del metodo NOTEARS con reti neurali. *arXiv preprint arXiv:1904.10098* (versione aggiornata). [Preprint — fonte del metodo del Lagrangiano aumentato per il vincolo di aciclicità in CV-JEPA]

Zhang, B. e Sennrich, R. (2019). Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 32. [Articolo — RMSNorm; formalizzato nella teoria Why3 dell'Appendice B]

Zheng, X., Aragam, B., Ravikumar, P. e Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 31. [Articolo — vincolo di aciclicità differenziabile NOTEARS $h(A) = \text{tr}(e^{\{A \circ A\}}) - d$; fonte primaria del termine T_7 nell'ELBO causale di CV-JEPA]

Nota sulle abbreviazioni adottate nelle citazioni nel corpo del testo

Le citazioni nel corpo della monografia adottano la forma (Autore anno) quando si fa riferimento a un'opera specifica e si citano nel testo le sole informazioni necessarie a identificare l'opera senza ambiguità. Laddove la stessa opera è invocata più volte nell'arco di pochi paragrafi, la citazione è abbreviata al solo cognome del primo autore dopo la prima occorrenza completa. Le citazioni ai capitoli della stessa monografia o delle due monografie gemelle del trittico adottano la forma «Capitolo N» o «seconda monografia, Capitolo N» senza ulteriori riferimenti bibliografici, data la natura cumulativa del progetto di ricerca che le tre opere costruiscono congiuntamente.